

This Page Is Inserted by IFW Operations
and is not a part of the Official Record

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

**As rescanning documents *will not* correct images,
please do not report the images to the
Image Problem Mailbox.**

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

DECLARATION OF TOD BEDILION, Ph.D.
UNDER 37 C.F.R. § 1.132

I, TOD BEDILION, Ph.D., declare and state as follows:

1. In April, 1996, I became the first employee of Synteni, Inc., where I served as Research Director until its acquisition by Incyte Corporation in early 1998. After Synteni's acquisition, I continued in the position of Director of Corporate Development at Incyte until May 11, 2001. I am currently the Director of Business Development at Genomic Health, Inc., Redwood City, California and an occasional Consultant to Incyte.

2. Synteni was founded to commercialize expression microarrays, microarrays in which expressed nucleic acids -- full-length cDNAs, fragments of full-length cDNAs, expressed sequence tags (ESTs) -- are arrayed on a common support to permit highly parallel detection and measurement of the expression of their cognate genes in a biological sample.

3. During my employ at Synteni, virtually all (if not all) of my work efforts were directed to the further technical development and the commercial exploitation of that microarray technology; given the small size of our shop, most of us had both technical and commercial responsibilities. The customer accounts for which I was personally responsible included large pharmaceutical companies, such as SmithKline

Beecham, large biotechnology companies, such as Genentech, and small research institutes, such as DNAX Inc.

4. From my very first interaction with our customers, consistently through to Synteni's acquisition by Incyte, I heard uniform, consistent, and emphatic requests that more genes be added to the arrays. This was true with respect to both our original microarrays, based on customer-provided genes and libraries, and our later, "generic", gene expression microarrays, based upon the unigene clone collection (our so-called "UniGem" arrays). From day 1, the pressure on us was to print ever more spots on the array. It was never a question: our customers wanted ever more genes on the array, each new gene-specific probe providing incrementally more value to the customer.¹

5. As a commercial enterprise, providing value to our customers was our major concern. Thus, to increase the value of our products and services in the marketplace -- to increase our ability to sell our microarrays and microarray services, their "salability" -- our efforts from the very beginning were devoted to increasing the number of specific genes whose expression could be detected with our microarrays.

6. Indeed, one of our major competitive advantages in the marketplace -- not just as regards other commercial suppliers, but also with respect to the innumerable laboratories and companies that were attempting to spot arrays in their own "home-brew" facilities -- was the number of

¹ I should note the customers were not asking for addition of probes specific to only those genes for which the biological function of the encoded gene product was known, but were asking for probes specific to any and all expressed genes.

distinct gene-specific probes that we provided on our expression microarrays. Our first 10,000 element UniGem array put the holy grail of gene expression analysis -- the human whole genome array -- within sight for the very first time (with respect to timing of the UniGEM program we began project planning and technology development in mid 1996 and delivered our first 10,000 element standard content human arrays in the first months of 1997 as I recall).

7. By the end of 1997, our efforts to provide the most comprehensive, and thus most valuable, human gene expression microarrays had been sufficiently successful that Incyte agreed to acquire Synteni for a reported \$80 million.

8. I declare further that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true, and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and may jeopardize the validity of any patent application in which this declaration is filed or any patent that issues thereon.



Tod Bedilion, Ph.D.

10/20/03

Date

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

DECLARATION OF VISHWANATH R. IYER, Ph.D.
UNDER 37 C.F.R. § 1.132

I, VISHWANATH R. IYER, Ph.D., declare and state as follows:

1. I am an Assistant Professor in the Section of Molecular Genetics and Microbiology, Institute of Cellular and Molecular Biology, University of Texas at Austin, where my laboratory currently studies global transcriptional control in yeast, gene expression programs during human cell proliferation, and genome-wide transcription factor targets in yeast and human. Immediately prior to this position, I spent four years as a postdoctoral fellow in the laboratory of Patrick O. Brown at Stanford University studying the transcriptional programs of yeast and of human cells. My curriculum vitae is attached hereto as Exhibit A.

2. Beginning in Dr. Brown's laboratory, where I helped to develop the first whole genome arrays for yeast and early versions of highly representative cDNA arrays for human cells, and continuing to the present day, I have used microarray-based gene expression analysis as a principal approach in much of my research.

3. Representative publications describing this work include:

DeRisi J. et al., "Exploring the metabolic and genetic control of gene expression on a genomic scale," *Science* 278:680-686 (1997);¹

Marton et al., "Drug target validation and identification of secondary drug target effects using DNA microarrays," *Nature Med.* 4:1293-1301 (1998);²

Iyer et al., "The transcriptional program in the response of human fibroblasts to serum," *Science* 283:83-87 (1999);³ and

Ross et al., "Systematic variation in gene expression patterns in human cancer cell lines," *Nature Genetics* 24: 227-235 (2000).⁴

Two of the papers describe our use of microarray-based expression profiling to explore the metabolic reprogramming that occurs during major environmental changes, both in yeast (DeRisi et al., during the shift from fermentation to respiration) and in human cells (Iyer et al., human fibroblasts exposed to serum). One reference describes our use of expression profile analysis in drug target validation and identification of secondary drug effects (Marton et al.). And one describes our use of expression profiling as a molecular phenotyping tool to discriminate among human cancer cells (Ross et al.).

4. Whether used to elucidate basic physiological responses, to study primary and secondary drug effects, or to discriminate and classify human cancers, expression profiling

¹ Attached hereto as Exhibit B.

² Attached hereto as Exhibit C.

³ Attached hereto as Exhibit D.

⁴ Attached hereto as Exhibit E.

as we have practiced it relies for its power on comparison of patterns of expression.

5. For example, we have demonstrated that we can use the presence or absence of a characteristic drug "signature" pattern of altered gene expression in drug-treated cells to explore the mechanism of drug action, and to identify secondary effects that can signal potentially deleterious drug side effects. As another example, we have demonstrated that gene expression patterns can be used to classify human tumor cell lines. While it is of course advantageous to know the biological function of the encoded gene products in order to reach a better understanding of the cellular mechanisms underlying these results, these pattern-based analyses do not require knowledge of the biological function of the encoded proteins.

6. The resolution of the patterns used in such comparisons is determined by the number of genes detected: the greater the number of genes detected, the higher the resolution of the pattern. It goes without saying that higher resolution patterns are generally more useful in such comparisons than lower resolution patterns. With such higher resolutions comes a correspondingly higher degree of statistical confidence for distinguishing different patterns, as well as identifying similar ones.

7. Each gene included as a probe on a microarray provides a signal that is specific to the cognate transcript, at least to a first approximation.⁵ Each new gene-specific

⁵ In a more nuanced view, it is certainly possible for a probe to signal the presence of a variety of splice variants of a single gene,

(Continued...)

probe added to a microarray thus increases the number of genes detectable by the device, increasing the resolving power of the device. As I note above, higher resolution patterns are generally more useful in comparisons than lower resolution patterns. Accordingly, each new gene probe added to a microarray increases the usefulness of the device in gene expression profiling analyses. This proposition is so well-established as to be virtually an axiom in the art, and has been as long as I have been working in the field, and certainly since the time I embarked on the production of whole genome arrays in early 1996. Simply put, arrays with fewer gene-specific probes are inferior to arrays with more gene-specific probes.

8. For example, our ability to subdivide cancers into discriminable classes by expression profiling is limited by the resolution of the patterns produced. With more genes contributing to the expression patterns, we can potentially draw finer distinctions among the patterns, thus subdividing otherwise indistinguishable cancers into a greater number of classes; the greater the number of classes, the greater the likelihood that the cancers classified together will respond similarly to therapeutic intervention, permitting better individualization of therapy and, we hope, better treatment outcomes.

9. If a gene does not change expression in an experiment, or if a gene is not expressed and produces no

(...Continued)

without discriminating among them, and for a probe to signal the presence of a variety of allelic variants of a single gene, again without discriminating among them.

signal in an experiment, that is not to say that the probe lacks usefulness on the array; it only means that an insufficient number of conditions have been sampled to identify expression changes. In fact, an experiment showing that a gene is not expressed or that its expression level does not change can be equally informative. To provide maximum versatility as a research tool, the microarray should include -- and as a biologist I would want my microarray to include -- each newly identified gene as a probe.

10. I declare further that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true, and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and may jeopardize the validity of any patent application in which this declaration is filed or any patent that issues thereon.



VISHWANATH R. IYER, Ph.D.

October 20, 2003

Date

Vishwanath R. Iyer

Assistant Professor

Section of Molecular Genetics and Microbiology
Institute of Cellular and Molecular Biology
MBB 3.212A, University of Texas at Austin
Austin, TX 78712-0159

Phone: 512-232-7833
Fax: 512-232-3432
Email: vishy@mail.utexas.edu

Education/Training

Bombay University, Mumbai, India	B.Sc. (1987), Chemistry & Biochemistry
M. S. University of Baroda, Baroda, India	M.Sc. (1989), Biotechnology
Harvard University, Cambridge MA	Ph.D. (1996), Genetics
Stanford University, Stanford CA	Post-doctoral (1996-2000), Genomics

Research Experience

- 9/00-5/03 Assistant professor, Section of Molecular Genetics and Microbiology, University of Texas, Austin TX
- Global transcriptional control in yeast
 - Gene expression programs during human cell proliferation
 - Genome-wide transcription factor targets in yeast and human
 - Collaborative microarray facility
- 5/96-8/00 Post-doctoral fellow Stanford University, Stanford CA
(Advisor: Dr. Patrick O. Brown)
- Yeast whole-genome ORF and intergenic microarrays
 - Human cDNA microarrays for expression profiling
- 9/89-4/96 Graduate student Harvard University, Cambridge MA
(Advisor: Dr. Kevin Struhl)
- Yeast transcriptional regulation

Honours and Awards

Government of India Biotechnology Fellowship (1987-1989)
University Grants Commission Junior Research Fellowship (1989)
Stanford University/NHGRI Genome Training Grant (1996)

Invited Conference talks (selected)

Invited Lecturer, NEC-Princeton Lectures in Biophysics
Princeton, NJ (June 1998)
Plenary Session Speaker, HGM '99 (HUGO Human Genome Meeting)
Brisbane, Australia (April 1999)
Invited Speaker, Gordon Research Conference "Human Molecular Genetics"
Newport, RI (August 2001)

Invited Speaker, Nature Genetics "Oncogenomics 2002" Conference
Dublin, Ireland (May 2002)
Invited Speaker, "Pathology Bioinformatics" Symposium, University of Michigan,
Ann Arbor, MI (November 2002)
Invited Speaker, "Systems Biology: Genomic Approaches to Transcriptional
Regulation" Cold Spring Harbor Laboratory Meeting (March 2003)
Symposium co-Chair and Speaker "Functional Genomics" American Society for
Biochemistry and Molecular Biology Meeting, San Diego, CA (April 2003)
Invited Speaker in Functional Genomics (Gene Networks) Symposium, International
Congress of Genetics, Melbourne Australia July 6-11 2003
Invited Speaker "BioArrays Europe 2003"
Cambridge, UK (Sep/Oct 2003)

Departmental Seminars

Texas A&M University Genetics and Biochemistry & Biophysics Departments,
October 24 2002
New York University School of Medicine, Department of Biochemistry,
November 20 2002
UT Southwestern Medical Center, Human Genetics Seminar Series,
May 5 2002
UCLA School of Medicine, Department of Human Genetics
June 2 2003
National Human Genome Research Institute
June 12 2003
Sanger Institute of the Wellcome Trust, Hinxton, UK
Sep 2003

Other Professional Activities

Reviewer for *Genome Biology*, *Genome Research*, *Nature Genetics*, *Science* (1998-
2003)
Instructor, Cold Spring Harbor Summer Course "Making and using DNA Microarrays"
(2000 - 2003)
Member, NIDDK Special Emphasis Review Panel ZDK1 (2001-2002)

Publications

1. Iyer V. & Struhl, K. (1995) Poly(dA:dT), a ubiquitous promoter element that stimulates transcription via its intrinsic DNA structure, *EMBO J.* 14: 2570-2579.
2. Iyer V. & Struhl, K. (1995) Mechanism of differential utilization of the his3 TR and TC TATA elements, *Mol. Cell. Biol.* 15: 7059-7066.
3. Iyer V. & Struhl K. (1996) Absolute mRNA levels and transcription initiation rates in *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. (USA)* 93:5208-5212.

4. DeRisi J. L., Iyer V. R. & Brown P. O. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278:680-686
5. Marton M. J., DeRisi J. L., Bennett H. A., Iyer V. R., Meyer M. R., Roberts C. J., Stoughton R., Burchard J., Slade D., Dai H., Bassett D. E. Jr., Hartwell L. H., Brown P. O. & Friend S. H. (1998) Drug target validation and identification of secondary drug target effects using DNA microarrays. *Nature Med.* 4:1293-1301
6. Lutfiyya L. L., Iyer V. R., DeRisi J., DeVit M. J., Brown P. O. & Johnston M. (1998) Characterization of three related glucose repressors and genes they regulate in *Saccharomyces cerevisiae*. *Genetics* 150:1377-1391
7. Spellman P. T., Sherlock G., Zhang M. Q., Iyer V. R., Anders K., Eisen M. B., Brown P. O., Botstein D. & Futcher B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell* 9:3273-3297
8. Iyer V. R., Eisen M. B., Ross D. T., Schuler G., Moore T., Lee J. C., F., Trent J. M., Staudt L. M., Hudson Jr. J., Boguski M. S., Lashkari D., Shalon D., Botstein D. & Brown P. O. (1999) The transcriptional program in the response of human fibroblasts to serum. *Science* 283:83-87
9. DeRisi J. L. & Iyer V. R. (1999) Genomics and array technology. *Curr. Opin. Oncol.* 11:76-79
10. Ross D. T., Scherf U., Eisen M. B., Perou C. M., Spellman P., Iyer V. R., Rees C., Jeffrey S. S., Van de Rijn M., Waltham M., Pergamenschikov A., Lee J. C. F., Lashkari D., Shalon D., Myers T. G., Weinstein J. N., Botstein D., & Brown P. O. (2000) Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genetics* 24: 227-235
11. Sudarsanam P., Iyer V. R., Brown P. O. & Winston F. (2000) Whole-genome expression analysis of *snf/swi* mutants of *S. cerevisiae*. *Proc. Natl. Acad. Sci. (USA)* 97: 3364-3369
12. Tran H. G., Steger D. J., Iyer V. R., & Johnson A. D. (2000) The chromo domain protein Chd1p from budding yeast is an ATP-dependent chromatin-modifying factor *EMBO J* 19: 2323-2331
13. Gross C., Kelleher M., Iyer V. R., Brown P. O., & Winge D. R.. (2000) Identification of the copper regulon in *Saccharomyces cerevisiae* by DNA microarrays. *J. Biol. Chem.* 275: 32310-32316
14. Reid J. L., Iyer V. R., Brown P. O. & Struhl K. (2000) Coordinate regulation of yeast ribosomal protein genes is associated with targeted recruitment of Esa1 histone acetylase. *Mol. Cell* 6: 1297-1307

15. Iyer V. R., Horak C., Scafe C. S., Botstein D., Snyder M. & Brown P. O. (2001) Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF *Nature* 409: 533-538
16. Miki R., Kadota K., Bono H., Mizuno Y., Tomaru Y., Carninci P., Itoh M., Shibata K., Kawai J., Konno H., Watanabe S., Sato K., Tokusumi Y., Kikuchi N., Ishii Y., Hamaguchi Y., Nishizuka I., Goto H., Nitanda H., Satomi S., Yoshiki A., Kusakabe M., DeRisi J.L., Eisen M.B., Iyer V.R., Brown P.O., Muramatsu M., Shimada H., Okazaki Y. & Hayashizaki Y. (2001) Delineating developmental and metabolic pathways in vivo by expression profiling using the RIKEN set of 18,816 full-length enriched mouse cDNA arrays *Proc. Natl. Acad. Sci. (USA)* 98: 2199-2204
17. Pollack J. R. & Iyer V.R. (2002) Characterizing the physical genome. *Nature Genetics* 32 suppl: 515-521
18. Iyer V. R. Microarray-based detection of DNA protein interactions: Chromatin Immunoprecipitation on Microarrays, in *DNA Microarrays: A Molecular Cloning Manual* (eds. Bowtell, D. & Sambrook, J.) 453-463 (Cold Spring Harbor Laboratory Press, 2003).
*(not peer reviewed)
19. Killion, P., Sherlock G. and Iyer V. R. (2003) The Longhorn Array Database, an open-source implementation of the Stanford Microarray Database *BMC Bioinformatics* 4: 32
20. Hahn J. S., Hu Z., Thiele D. J. & Iyer V. R. Genome-Wide Analysis of the Biology of Stress Responses Through Heat Shock Transcription Factor (submitted to *PNAS*)
21. Kim J. & Iyer V.R. The global role of TBP recruitment to promoters in mediating gene expression profiles (manuscript in preparation)

Current/Pending Research Support

U01 AA13518-01 Adron Harris (PI) 25% effort

9/28/01 - 9/27/06

NIH/NIAAA

"INIA: Microarray Core"

This proposal was a response to the Integrative Neuroscience Initiative on Alcoholism (INIA) RFA-AA-01-002. The overall goal is to support the use of microarray technology to define changes in gene expression that either predict or accompany excessive alcohol consumption.

Role: Co-investigator

003658-0223-2001 Iyer (PI) 16% effort

01/01/02 - 08/31/04

Texas Higher Education Coordinating Board (ARP)

"Microarray based global mapping of DNA-protein interactions at promoters in human cells"

This is a pilot project to map the in vivo interactions of transcription factors with human promoters

Role: PI

Information Technology Research 0325116 R. Mooney (PI) 9% effort

09/01/03 - 08/31/07

NSF

"Feedback from Multi-Source Data Mining to Experimentation for Gene Network Discovery"

Role: Co-investigator

1 R01 CA95548-01A2 (pending) Iyer (PI) 25% effort

12/1/03 - 11/30/08

NIH

"Analysis of genome-wide transcriptional control in yeast"

This is a project to identify stress responsive transcription factor targets in yeast through the use of DNA microarrays

Role: PI

Breast Cancer Idea Award (pending) Iyer (PI) 10% effort

1/1/04 - 12/31/06

US Army Medical Research and Materiel Command

"Genome-wide chromosomal targets of oncogenic transcription factors"

This is a project aimed at identifying direct chromosomal targets of c-myc and ER in human cells through the use of a novel sequence tag analysis method.

Role: PI

003658-0531-2003 (pending) Marcotte (PI) 8% effort

01/01/04 - 12/31/05

Texas Higher Education Coordinating Board (ATP)

"Cell arrays: A novel high-throughput platform for measuring gene function on a genomic scale"

This proposal is aimed at developing a novel microarray based platform for automated, high-throughput microscopic imaging of cells, allowing rapid and systematic evaluation of gene function.

- Fischer-Vize, *Science* 270, 1828 (1995).
35. T. C. James and S. C. Elgin, *Mol. Cell Biol.* 6, 3862 (1986); R. Paro and D. S. Hogness, *Proc. Natl. Acad. Sci. U.S.A.* 88, 263 (1991); B. Tschiersch et al., *EMBO J.* 13, 3822 (1994); M. T. Madiredi et al., *Cell* 87, 75 (1996); D. G. Stokes, K. D. Tartof, R. P. Perry, *Proc. Natl. Acad. Sci. U.S.A.* 93, 7137 (1996).
36. P. M. Palosaari et al., *J. Biol. Chem.* 266, 10750 (1991); A. Schmitz, K. H. Gartemann, J. Fiedler, E. Grund, R. Eichenlaub, *Appl. Environ. Microbiol.* 58, 4068 (1992); V. Sharma, K. Suvama, R. Megathanan, M. E. Hudspeth, *J. Bacteriol.* 174, 5057 (1992); M. Kanazawa et al., *Enzyme Protein* 47, 9 (1993); Z. L. Boynton, G. N. Bennet, F. B. Rudolph, *J. Bacteriol.* 178, 3015 (1996).
37. M. Ho et al., *Cell* 77, 869 (1994).
38. W. Hendriks et al., *J. Cell Biochem.* 59, 418 (1995).
39. We thank H. Skaletsky and F. Lewitter for help with

Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale

Joseph L. DeRisi, Vishwanath R. Iyer, Patrick O. Brown*

DNA microarrays containing virtually every gene of *Saccharomyces cerevisiae* were used to carry out a comprehensive investigation of the temporal program of gene expression accompanying the metabolic shift from fermentation to respiration. The expression profiles observed for genes with known metabolic functions pointed to features of the metabolic reprogramming that occur during the diauxic shift, and the expression patterns of many previously uncharacterized genes provided clues to their possible functions. The same DNA microarrays were also used to identify genes whose expression was affected by deletion of the transcriptional co-repressor *TUP1* or overexpression of the transcriptional activator *YAP1*. These results demonstrate the feasibility and utility of this approach to genomewide exploration of gene expression patterns.

using a simple robotic printing device (9). Cells from an exponentially growing culture of yeast were inoculated into fresh medium and grown at 30°C for 21 hours. After an initial 9 hours of growth, samples were harvested at seven successive 2-hour intervals, and mRNA was isolated (10). Fluorescently labeled cDNA was prepared by reverse transcription in the presence of Cy3(green)- or Cy5(red)-labeled deoxyuridine triphosphate (dUTP) (11) and then hybridized to the microarrays (12). To maximize the reliability with which changes in expression levels could be discerned, we labeled cDNA prepared from cells at each successive time point with Cy5, then mixed it with a Cy3-labeled "reference" cDNA sample prepared from cells harvested at the first interval after inoculation. In this experimental design, the relative fluorescence intensity measured for the Cy3 and Cy5 fluorors at each array element provides a reliable measure of the relative abundance of the corresponding mRNA in the two cell populations (Fig. 1). Data from the series of seven samples (Fig. 2), consisting of more than 43,000 expression-ratio measurements, were organized into a database to facilitate efficient exploration and analysis of the results. This database is publicly available on the Internet (13).

During exponential growth in glucose-rich medium, the global pattern of gene expression was remarkably stable. Indeed, when gene expression patterns between the first two cell samples (harvested at a 2-hour interval) were compared, mRNA levels differed by a factor of 2 or more for only 19 genes (0.3%), and the largest of these differences was only 2.7-fold (14). However, as glucose was progressively depleted from the growth media during the course of the experiment, a marked change was seen in the global pattern of gene expression. mRNA levels for approximately 710 genes were induced by a factor of at least 2, and the mRNA levels for approximately 1030 genes declined by a factor of at least 2. Messenger RNA levels for 183 genes increased by a factor of at least 4, and mRNA levels for 203 genes diminished by a factor of at least 4. About half of these differentially expressed genes have no currently recognized function and are not yet named. Indeed, more than 400 of the differentially expressed genes have no apparent homology

The complete sequences of nearly a dozen microbial genomes are known, and in the next several years we expect to know the complete genome sequences of several metazoans, including the human genome. Defining the role of each gene in these genomes will be a formidable task, and understanding how the genome functions as a whole in the complex natural history of a living organism presents an even greater challenge.

Knowing when and where a gene is expressed often provides a strong clue as to its biological role. Conversely, the pattern of genes expressed in a cell can provide detailed information about its state. Although regulation of protein abundance in a cell is by no means accomplished solely by regulation of mRNA, virtually all differences in cell type or state are correlated with changes in the mRNA levels of many genes. This is fortuitous because the only specific reagent required to measure the abundance of the mRNA for a specific gene is a cDNA sequence. DNA microarrays, consisting of thousands of individual gene sequences printed in a high-density array on a glass microscope slide (1, 2), provide a practical and economical tool for studying gene expression on a very large scale (3-6).

Saccharomyces cerevisiae is an especially

favorable organism in which to conduct a systematic investigation of gene expression. The genes are easy to recognize in the genome sequence, cis regulatory elements are generally compact and close to the transcription units, much is already known about its genetic regulatory mechanisms, and a powerful set of tools is available for its analysis.

A recurring cycle in the natural history of yeast involves a shift from anaerobic (fermentation) to aerobic (respiration) metabolism. Inoculation of yeast into a medium rich in sugar is followed by rapid growth fueled by fermentation, with the production of ethanol. When the fermentable sugar is exhausted, the yeast cells turn to ethanol as a carbon source for aerobic growth. This switch from anaerobic growth to aerobic respiration upon depletion of glucose, referred to as the diauxic shift, is correlated with widespread changes in the expression of genes involved in fundamental cellular processes such as carbon metabolism, protein synthesis, and carbohydrate storage (7). We used DNA microarrays to characterize the changes in gene expression that take place during this process for nearly the entire genome, and to investigate the genetic circuitry that regulates and executes this program.

Yeast open reading frames (ORFs) were amplified by the polymerase chain reaction (PCR), with a commercially available set of primer pairs (8). DNA microarrays, containing approximately 6400 distinct DNA sequences, were printed onto glass slides by

Department of Biochemistry, Stanford University School of Medicine, Howard Hughes Medical Institute, Stanford, CA 94305-5428, USA.

*To whom correspondence should be addressed. E-mail: pbrown@cmgm.stanford.edu

The global view of changes in expression of genes with known functions provides a vivid picture of the way in which the cell adapts to a changing environment. Figure 3 shows a portion of the yeast metabolic pathways involved in carbon and energy metabolism. Mapping the changes we observed in the mRNAs encoding each enzyme onto this framework allowed us to infer the redirection in the flow of metabolites through this system. We observed large inductions of the genes coding for the enzymes aldehyde dehydrogenase (*ALD2*) and acetyl-coenzyme A (CoA) synthase (*ACS1*), which function together to convert the products of alcohol dehydrogenase into acetyl-CoA, which in turn is used to fuel the tricarboxylic acid (TCA) cycle and the glyoxylate cycle. The concomitant shutdown of transcription of the genes encoding pyruvate decarboxylase and induction of pyruvate carboxylase rechannels pyruvate away from acetaldehyde, and instead to oxalacetate, where it can serve to supply the TCA cycle and gluconeogenesis. Induction of the pivotal genes *PCK1*, encoding phosphoenolpyruvate carboxykinase, and *FBP1*, encoding fructose 1,6-bisphosphatase, switches the directions of two key irreversible steps in glycolysis, reversing the flow of metabolites along the reversible steps of the glycolytic pathway toward the essential biosynthetic precursor, glucose-6-phosphate. Induction of the genes coding for the trehalose synthase and glycogen synthase complexes promotes channeling of glucose-6-phosphate into these carbohydrate storage pathways.

Just as the changes in expression of genes encoding pivotal enzymes can provide insight into metabolic reprogramming, the behavior of large groups of functionally related genes can provide a broad view of the systematic way in which the yeast cell adapts to a changing environment (Fig. 4). Several classes of genes, such as cytochrome c-related genes and those involved in the TCA/glyoxylate cycle and carbohydrate storage, were coordinately induced by glucose exhaustion. In contrast, genes devoted to protein synthesis, including ribosomal proteins, tRNA synthetases, and translation, elongation, and initiation factors, exhibited a coordinated decrease in expression. More than 95% of ribosomal genes showed at least twofold decreases in expression during the diauxic shift (Fig. 4) (13). A noteworthy and illuminating exception was that the

genes encoding mitochondrial ribosomal genes were generally induced rather than repressed after glucose limitation, highlighting the requirement for mitochondrial biogenesis (13). As more is learned about the functions of every gene in the yeast genome, the ability to gain insight into a cell's response to a changing environment through its global gene expression patterns will become increasingly powerful.

Several distinct temporal patterns of expression could be recognized, and sets of genes could be grouped on the basis of the similarities in their expression patterns. The characterized members of each of these groups also shared important similarities in their functions. Moreover, in most cases, common regulatory mechanisms could be inferred for sets of genes with similar expression profiles. For example, seven genes showed a late induction profile, with mRNA levels increasing by more than ninefold at

the last timepoint but less than threefold at the preceding timepoint (Fig. 5B). All of these genes were known to be glucose-repressed, and five of the seven were previously noted to share a common upstream activating sequence (UAS), the carbon source response element (CSRE) (16–20). A search in the promoter regions of the remaining two genes, *ACR1* and *IDP2*, revealed that *ACR1*, a gene essential for *ACS1* activity, also possessed a consensus CSRE motif, but interestingly, *IDP2* did not. A search of the entire yeast genome sequence for the consensus CSRE motif revealed only four additional candidate genes, none of which showed a similar induction.

Examples from additional groups of genes that shared expression profiles are illustrated in Fig. 5, C through F. The sequences upstream of the named genes in Fig. 5C all contain stress response elements (STRE), and with the exception

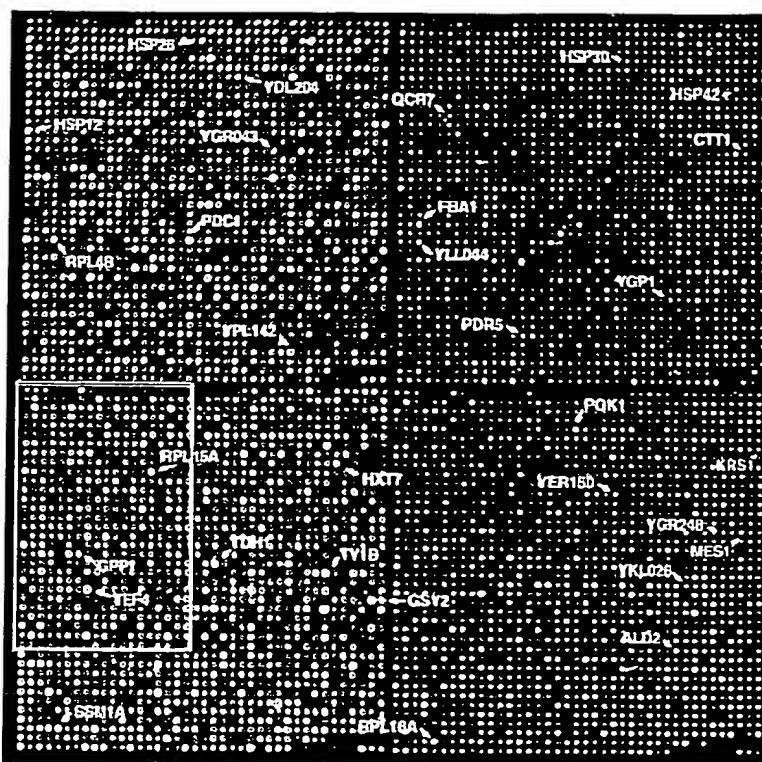


Fig. 1. Yeast genome microarray. The actual size of the microarray is 18 mm by 18 mm. The microarray was printed as described (9). This image was obtained with the same fluorescent scanning confocal microscope used to collect all the data we report (49). A fluorescently labeled cDNA probe was prepared from mRNA isolated from cells harvested shortly after inoculation (culture density of $<5 \times 10^6$ cells/ml and media glucose level of 19 g/liter) by reverse transcription in the presence of Cy3-dUTP. Similarly, a second probe was prepared from mRNA isolated from cells taken from the same culture 9.5 hours later (culture density of $\sim 2 \times 10^6$ cells/ml, with a glucose level of <0.2 g/liter) by reverse transcription in the presence of Cy5-dUTP. In this image, hybridization of the Cy3-dUTP-labeled cDNA (that is, mRNA expression at the initial timepoint) is represented as a green signal, and hybridization of Cy5-dUTP-labeled cDNA (that is, mRNA expression at 9.5 hours) is represented as a red signal. Thus, genes induced or repressed after the diauxic shift appear in this image as red and green spots, respectively. Genes expressed at roughly equal levels before and after the diauxic shift appear in this image as yellow spots.

of HSP42, have previously been shown to be controlled at least in part by these elements (21–24). Inspection of the sequences upstream of HSP42 and the two uncharacterized genes shown in Fig. 5C, YKL026c, a hypothetical protein with similarity to glutathione peroxidase, and YGR043c, a putative transaldolase, revealed that each of these genes also possess repeated upstream copies of the stress-responsive CCCCT motif. Of the 13 additional genes in the yeast genome that shared this expression profile [including HSP30, ALD2, OM45, and 10 uncharacterized ORFs (25)], nine contained one or more recognizable STRE sites in their upstream regions.

The heterotrimeric transcriptional activator complex HAP2,3,4 has been shown to be responsible for induction of several genes important for respiration (26–28). This complex binds a degenerate consensus sequence known as the CCAAT box (26). Computer analysis, using the consensus sequence TNRYTGGB (29), has suggested that a large number of genes involved in respiration may be specific targets of HAP2,3,4 (30). Indeed, a putative HAP2,3,4 binding site could be found in the sequences upstream of each of the seven cytochrome c-related genes that showed the greatest magnitude of induction (Fig. 5D). Of 12 additional cytochrome c-related genes that were induced, HAP2,3,4 binding sites were present in all but one. Significantly, we found that transcription of HAP4 itself was induced nearly ninefold concomitant with the diauxic shift.

Control of ribosomal protein biogenesis is mainly exerted at the transcriptional level, through the presence of a common upstream-activating element (UAS_{rp}) that is recognized by the Rap1 DNA-binding protein (31, 32). The expression profiles of seven ribosomal proteins are shown in Fig. 5F. A search of the sequences upstream of all seven genes revealed consensus Rap1-binding motifs (33). It has been suggested that declining Rap1 levels in the cell during starvation may be responsible for the decline in ribosomal protein gene expression (34). Indeed, we observed that the abundance of RAP1 mRNA diminished by 4.4-fold, at about the time of glucose exhaustion.

Of the 149 genes that encode known or putative transcription factors, only two, HAP4 and SIP4, were induced by a factor of more than threefold at the diauxic shift. SIP4 encodes a DNA-binding transcriptional activator that has been shown to interact with Snf1, the "master regulator" of glucose repression (35). The eightfold induction of SIP4 upon depletion of glucose strongly suggests a role in the induction of

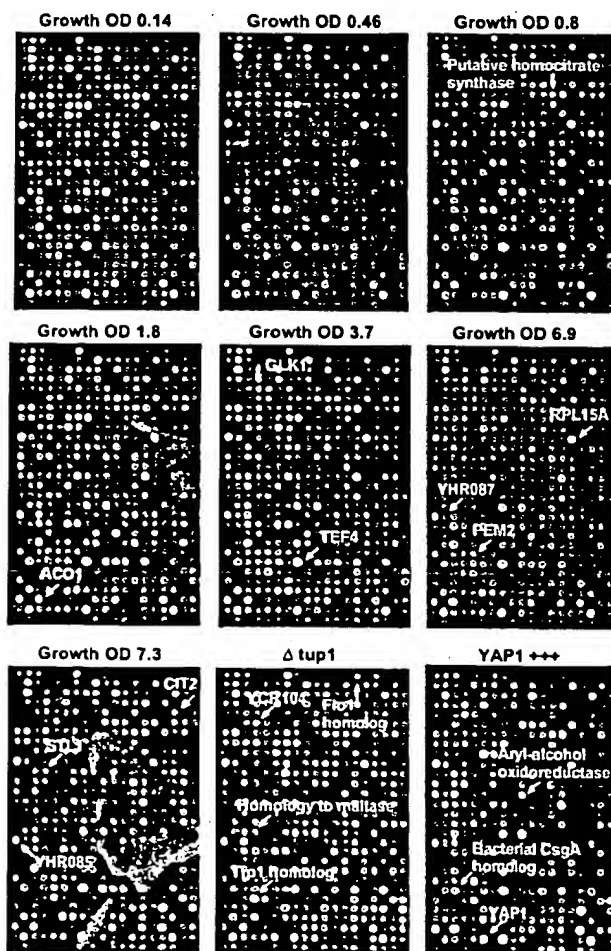
downstream genes at the diauxic shift.

Although most of the transcriptional responses that we observed were not previously known, the responses of many genes during the diauxic shift have been described. Comparison of the results we obtained by DNA microarray hybridization with previously reported results therefore provided a strong test of the sensitivity and accuracy of this approach. The expression patterns we observed for previously characterized genes showed almost perfect concordance with previously published results (36). Moreover, the differential expression measurements obtained by DNA microarray hybridization were reproducible in duplicate experiments. For example, the remarkable changes in gene expression between cells harvested immediately after inoculation and immediately after the diauxic shift (the first and sixth intervals in this time series) were measured in duplicate, independent DNA microarray hybridizations. The correlation coefficient for two complete sets of expression ratio measurements was 0.87, and for more than 95% of the genes, the expres-

sion ratios measured in these duplicate experiments differed by less than a factor of 2. However, in a few cases, there were discrepancies between our results and previous results, pointing to technical limitations that will need to be addressed as DNA microarray technology advances (37, 38). Despite the noted exceptions, the high concordance between the results we obtained in these experiments and those of previous studies provides confidence in the reliability and thoroughness of the survey.

The changes in gene expression during this diauxic shift are complex and involve integration of many kinds of information about the nutritional and metabolic state of the cell. The large number of genes whose expression is altered and the diversity of temporal expression profiles observed in this experiment highlight the challenge of understanding the underlying regulatory mechanisms. One approach to defining the contributions of individual regulatory genes to a complex program of this kind is to use DNA microarrays to identify genes whose expression is affected

Fig. 2. The section of the array indicated by the gray box in Fig. 1 is shown for each of the experiments described here. Representative genes are labeled. In each of the arrays used to analyze gene expression during the diauxic shift, red spots represent genes that were induced relative to the initial timepoint, and green spots represent genes that were repressed relative to the initial timepoint. In the arrays used to analyze the effects of the *tup1Δ* mutation and YAP1 overexpression, red spots represent genes whose expression was increased, and green spots represent genes whose expression was decreased by the genetic modification. Note that distinct sets of genes are induced and repressed in the different experiments. The complete images of each of these arrays can be viewed on the Internet (13). Cell density as measured by optical density (OD) at 600 nm was used to measure the growth of the culture.



by mutations in each putative regulatory gene. As a test of this strategy, we analyzed the genomewide changes in gene expression that result from deletion of the *TUP1* gene. Transcriptional repression of many genes by glucose requires the DNA-binding repressor

Mig1 and is mediated by recruiting the transcriptional co-repressors Tup1 and Cyc8/Ssn6 (39). Tup1 has also been implicated in repression of oxygen-regulated, mating-type-specific, and DNA-damage-inducible genes (40).

Wild-type yeast cells and cells bearing a deletion of the *TUP1* gene (*tup1Δ*) were grown in parallel cultures in rich medium containing glucose as the carbon source. Messenger RNA was isolated from exponentially growing cells from the two populations and used to prepare cDNA labeled with Cy3 (green) and Cy5 (red), respectively (11). The labeled probes were mixed and simultaneously hybridized to the microarray. Red spots on the microarray therefore represented genes whose transcription was induced in the *tup1Δ* strain, and thus presumably repressed by Tup1 (41). A representative section of the microarray (Fig. 2, bottom middle panel) illustrates that the genes whose expression was affected by the *tup1Δ* mutation, were, in general, distinct from those induced upon glucose exhaustion [complete images of all the arrays shown in Fig. 2 are available on the Internet (13)]. Nevertheless, 34 (10%) of the genes that were induced by a factor of at least 2 after the diauxic shift were similarly induced by deletion of *TUP1*, suggesting that these genes may be subject to *TUP1*-mediated repression by glucose. For example, *SUC2*, the gene encoding invertase, and all five hexose transporter genes that were induced during the course of the diauxic shift were similarly induced, in duplicate experiments, by the deletion of *TUP1*.

The set of genes affected by Tup1 in this experiment also included α -glucosidases, the mating-type-specific genes *MFA1* and *MFA2*, and the DNA damage-inducible *RNR2* and *RNR4*, as well as genes involved in flocculation and many genes of unknown function. The hybridization signal corresponding to expression of *TUP1* itself was also severely reduced because of the (incomplete) deletion of the transcription unit in the *tup1Δ* strain, providing a positive control in the experiment (42).

Many of the transcriptional targets of Tup1 fell into sets of genes with related biochemical functions. For instance, although only about 3% of all yeast genes appeared to be *TUP1*-repressed by a factor of more than 2 in duplicate experiments under these conditions, 6 of the 13 genes that have been implicated in flocculation (15) showed a reproducible increase in expression of at least twofold when *TUP1* was deleted. Another group of related genes that appeared to be subject to *TUP1* repression encodes the serine-rich cell wall mannoproteins, such as *Tipl* and *Tirl/Srp1* which are induced by cold shock and other stresses (43), and similar, serine-poor proteins, the seripauperins (44). Messenger RNA levels for 23 of the 26 genes in this group were reproducibly elevated by at least 2.5-fold in the *tup1Δ*

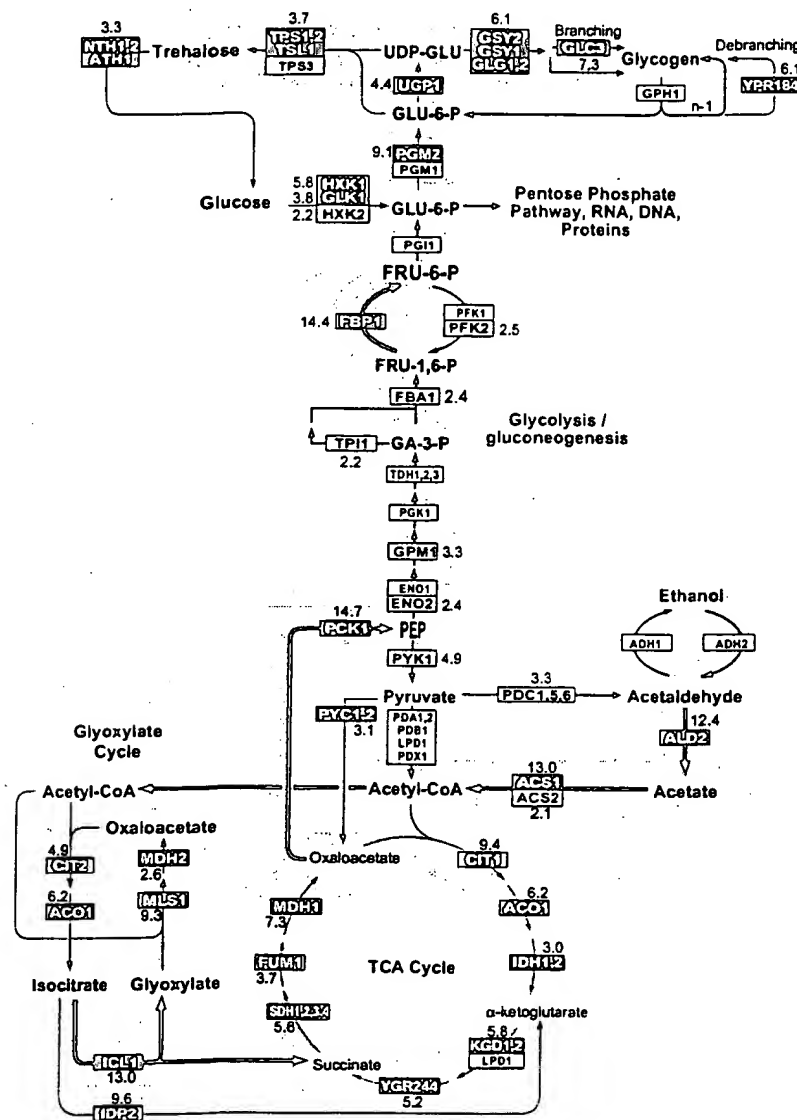


Fig. 3. Metabolic reprogramming inferred from global analysis of changes in gene expression. Only key metabolic intermediates are identified. The yeast genes encoding the enzymes that catalyze each step in this metabolic circuit are identified by name in the boxes. The genes encoding succinyl-CoA synthase and glycogen-debranching enzyme have not been explicitly identified, but the ORFs *YGR244* and *YPR184* show significant homology to known succinyl-CoA synthase and glycogen-debranching enzymes, respectively, and are therefore included in the corresponding steps in this figure. Red boxes with white lettering identify genes whose expression increases in the diauxic shift. Green boxes with dark green lettering identify genes whose expression diminishes in the diauxic shift. The magnitude of induction or repression is indicated for these genes. For multimeric enzyme complexes, such as succinate dehydrogenase, the indicated fold-induction represents an unweighted average of all the genes listed in the box. Black and white boxes indicate no significant differential expression (less than twofold). The direction of the arrows connecting reversible enzymatic steps indicate the direction of the flow of metabolic intermediates, inferred from the gene expression pattern, after the diauxic shift. Arrows representing steps catalyzed by genes whose expression was strongly induced are highlighted in red. The broad gray arrows represent major increases in the flow of metabolites after the diauxic shift, inferred from the indicated changes in gene expression.

strain, and 18 of these genes were induced by more than sevenfold when *TUP1* was deleted. In contrast, none of 83 genes that could be classified as putative regulators of the cell division cycle were induced more than twofold by deletion of *TUP1*. Thus, despite the diversity of the regulatory systems that employ Tup1, most of the genes that it regulates under these conditions fall into a limited number of distinct functional classes.

Because the microarray allows us to monitor expression of nearly every gene in yeast, we can, in principle, use this approach to identify all the transcriptional targets of a regulatory protein like Tup1. It is important to note, however, that in any single experiment of this kind we can only recognize those target genes that are normally repressed (or induced) under the conditions of the experiment. For instance, the experiment described here analyzed a MAT α strain in which *MFA1* and *MFA2*, the genes encoding the α -factor mating pheromone precursor, are normally repressed. In the isogenic *tup1 Δ* strain, these genes were inappropriately expressed, reflecting the role that Tup1 plays in their repression. Had we instead carried out this experiment with a MAT α strain (in which expression of *MFA1* and *MFA2* is not repressed), it would not have been possible to conclude anything regarding the role of Tup1 in the repression of these genes. Conversely, we cannot distinguish indirect effects of the chronic absence of Tup1 in the mutant strain from effects directly attributable to its participation in repressing the transcription of a gene.

Another simple route to modulating the activity of a regulatory factor is to overexpress the gene that encodes it. *YAP1* encodes a DNA-binding transcription factor belonging to the b-zip class of DNA-binding proteins. Overexpression of *YAP1* in yeast confers increased resistance to hydrogen peroxide, *o*-phenanthroline, heavy metals, and osmotic stress (45). We analyzed differential gene expression between a wild-type strain bearing a control plasmid and a strain with a plasmid expressing *YAP1* under the control of the strong *GAL1-10* promoter, both grown in galactose (that is, a condition that induces *YAP1* overexpression). Complementary DNA from the control and *YAP1* overexpressing strains, labeled with Cy3 and Cy5, respectively, was prepared from mRNA isolated from the two strains and hybridized to the microarray. Thus, red spots on the array represent genes that were induced in the strain overexpressing *YAP1*.

Of the 17 genes whose mRNA levels increased by more than threefold when

YAP1 was overexpressed in this way, five bear homology to aryl-alcohol oxidoreductases (Fig. 2 and Table 1). An additional four of the genes in this set also belong to the general class of dehydrogenases/oxidoreductases. Very little is known about the role of aryl-alcohol oxidoreductases in *S. cerevisiae*, but these enzymes have been isolated from ligninolytic fungi, in which they participate in coupled redox reactions, oxidizing aromatic, and aliphatic unsaturated alcohols to aldehydes with the production of hydrogen peroxide (46, 47). The fact that a remarkable fraction of the targets identified in this experiment belong to the same small, functional group of oxidoreductases suggests that these genes

might play an important protective role during oxidative stress. Transcription of a small number of genes was reduced in the strain overexpressing *Yap1*. Interestingly, many of these genes encode sugar permeases or enzymes involved in inositol metabolism.

We searched for *Yap1*-binding sites (TTACTAA or TGACTAA) in the sequences upstream of the target genes we identified (48). About two-thirds of the genes that were induced by more than threefold upon *Yap1* overexpression had one or more binding sites within 600 bases upstream of the start codon (Table 1), suggesting that they are directly regulated by *Yap1*. The absence of canonical *Yap1*-bind-

Fig. 4. Coordinated regulation of functionally related genes. The curves represent the average induction or repression ratios for all the genes in each indicated group. The total number of genes in each group was as follows: ribosomal proteins, 112; translation elongation and initiation factors, 25; tRNA synthetases (excluding mitochondrial synthetases), 17; glycogen and trehalose synthesis and degradation, 15; cytochrome c oxidase and reductase proteins, 19; and TCA- and glyoxylate-cycle enzymes, 24.

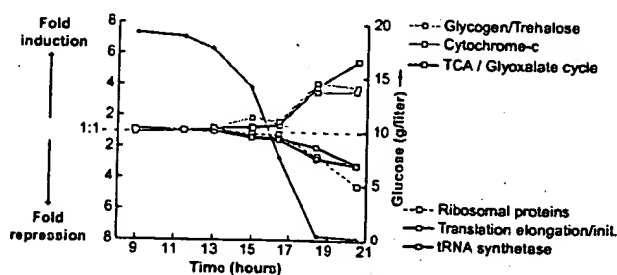


Table 1. Genes induced by *YAP1* overexpression. This list includes all the genes for which mRNA levels increased by more than twofold upon *YAP1* overexpression in both of two duplicate experiments, and for which the average increase in mRNA level in the two experiments was greater than threefold (50). Positions of the canonical *Yap1* binding sites upstream of the start codon, when present, and the average fold-increase in mRNA levels measured in the two experiments are indicated.

ORF	Distance of <i>Yap1</i> site from ATG	Gene	Description	Fold-increase
YNL331C	162-222 (5 sites)	<i>YAP1</i>	Putative aryl-alcohol reductase	12.9
YKL071W			Similarity to bacterial <i>csgA</i> protein	10.4
YML007W			Transcriptional activator involved in oxidative stress response	9.8
YFL056C	223, 242		Homology to aryl-alcohol dehydrogenases	9.0
YLL060C	98		Putative glutathione transferase	7.4
YOL165C	266		Putative aryl-alcohol dehydrogenase (NADP+)	7.0
YCR107W	409	<i>ATR1</i>	Putative aryl-alcohol reductase	6.5
YML116W			Aminotriazole and 4-nitroquinoline resistance protein	6.5
YBR008C	142, 167, 364		Homology to benomyl/methotrexate resistance protein	6.1
YCLX08C	148, 212	<i>OYE3</i>	Hypothetical protein	6.1
YJR155W			Putative aryl-alcohol dehydrogenase	6.0
YPL171C			NAPDH dehydrogenase (old yellow enzyme), isoform 3	5.8
YLR460C	167, 317		Homology to hypothetical proteins YCR102c and YNL134c	4.7
YKR076W	178		Homology to hypothetical protein YMR251w	4.5
YHR179W	327	<i>OYE2</i>	NAD(P)H oxidoreductase (old yellow enzyme), isoform 1	4.1
YML131W	507		Similarity to <i>A. thaliana</i> zeta-crystallin homolog	3.7
YOL126C		<i>MDH2</i>	Malate dehydrogenase	3.3

ing sites upstream of the others may reflect an ability of Yap1 to bind sites that differ from the canonical binding sites, perhaps in cooperation with other factors, or less likely, may represent an indirect effect of Yap1 overexpression, mediated by one or more intermediary factors. Yap1 sites were found only four times in the corresponding region of an arbitrary set of 30 genes that were not differentially regulated by Yap1.

Use of a DNA microarray to characterize the transcriptional consequences of mutations affecting the activity of regulatory molecules provides a simple and powerful approach to dissection and characterization of regulatory pathways and net-

works. This strategy also has an important practical application in drug screening. Mutations in specific genes encoding candidate drug targets can serve as surrogates for the ideal chemical inhibitor or modulator of their activity. DNA microarrays can be used to define the resulting signature pattern of alterations in gene expression, and then subsequently used in an assay to screen for compounds that reproduce the desired signature pattern.

DNA microarrays provide a simple and economical way to explore gene expression patterns on a genomic scale. The hurdles to extending this approach to any other organism are minor. The equipment

required for fabricating and using DNA microarrays (9) consists of components that were chosen for their modest cost and simplicity. It was feasible for a small group to accomplish the amplification of more than 6000 genes in about 4 months and, once the amplified gene sequences were in hand, only 2 days were required to print a set of 110 microarrays of 6400 elements each. Probe preparation, hybridization, and fluorescent imaging are also simple procedures. Even conceptually simple experiments, as we described here, can yield vast amounts of information. The value of the information from each experiment of this kind will progressively increase as more is learned about the functions of each gene and as additional experiments define the global changes in gene expression in diverse other natural processes and genetic perturbations. Perhaps the greatest challenge now is to develop efficient methods for organizing, distributing, interpreting, and extracting insights from the large volumes of data these experiments will provide.

REFERENCES AND NOTES

1. M. Schena, D. Shalon, R. W. Davis, P. O. Brown, *Science* 270, 467 (1995).
2. D. Shalon, S. J. Smith, P. O. Brown, *Genome Res.* 6, 639 (1996).
3. D. Lashkari, *Proc. Natl. Acad. Sci. U.S.A.*, in press.
4. J. DeRisi et al., *Nature Genet.* 14, 457 (1996).
5. D. J. Lockhart et al., *Nature Biotechnol.* 14, 1675 (1996).
6. M. Chee et al., *Science* 274, 610 (1996).
7. M. Johnston and M. Carlson, in *The Molecular Biology of the Yeast Saccharomyces: Gene Expression*, E. W. Jones, J. R. Pringle, J. R. Broach, Eds. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1992), p. 193.
8. Primers for each known or predicted protein coding sequence were supplied by Research Genetics. PCR was performed with the protocol supplied by Research Genetics, using genomic DNA from yeast strain S288C as a template. Each PCR product was verified by agarose gel electrophoresis and was deemed correct if the lane contained a single band of appropriate mobility. Failures were marked as such in the database. The overall success rate for a single-pass amplification of 6116 ORFs was ~94.5%.
9. Glass slides (Gold Seal) were cleaned for 2 hours in a solution of 2 N NaOH and 70% ethanol. After rinsing in distilled water, the slides were then treated with a 1:5 dilution of poly-L-lysine adhesive solution (Sigma) for 1 hour, and then dried for 5 min at 40°C in a vacuum oven. DNA samples from 100- μ l PCR reactions were purified by ethanol purification in 96-well microtiter plates. The resulting precipitates were resuspended in 3 \times standard saline citrate (SSC) and transferred to new plates for arraying. A custom-built arraying robot was used to print on a batch of 110 slides. Details of the design of the microarrayer are available at cmgm.stanford.edu/pbrown. After printing, the microarrays were rehydrated for 30 s in a humid chamber and then snap-dried for 2 s on a hot plate (100°C). The DNA was then ultraviolet (UV)-crosslinked to the surface by subjecting the slides to 60 mJ of energy (Stratagene Stratalinker). The rest of the poly-L-lysine surface was blocked by a 15-min incubation in a solution of 70 mM succinic anhydride dissolved in a solution consisting of 315 ml of 1-methyl-2-pyrrolidinone (Aldrich) and 35 ml of 1 M boric acid (pH 8.0). Directly after the blocking reac-

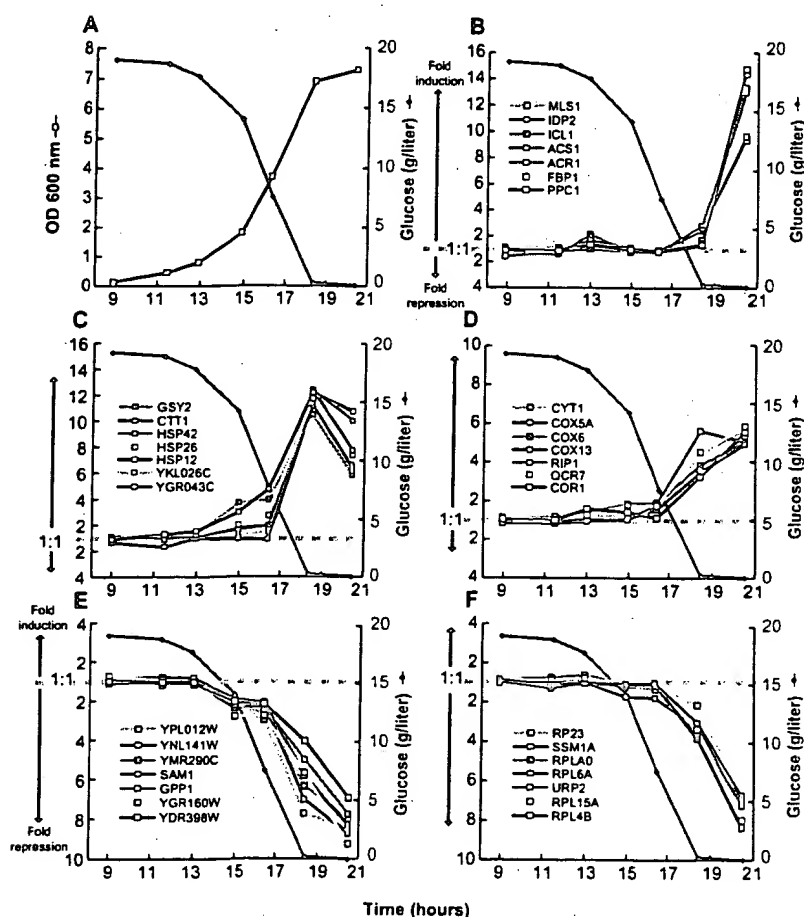


Fig. 5. Distinct temporal patterns of induction or repression help to group genes that share regulatory properties. (A) Temporal profile of the cell density, as measured by OD at 600 nm and glucose concentration in the media. (B) Seven genes exhibited a strong induction (greater than ninefold) only at the last timepoint (20.5 hours). With the exception of *IDP2*, each of these genes has a CSRE UAS. There were no additional genes observed to match this profile. (C) Seven members of a class of genes marked by early induction with a peak in mRNA levels at 18.5 hours. Each of these genes contains STRE motif repeats in their upstream promoter regions. (D) Cytochrome c oxidase and ubiquinol cytochrome c reductase genes. Marked by an induction coincident with the diauxic shift, each of these genes contains a consensus binding motif for the HAP2,3,4 protein complex. At least 17 genes shared a similar expression profile. (E) *SAM1*, *GPP1*, and several genes of unknown function are repressed before the diauxic shift, and continue to be repressed upon entry into stationary phase. (F) Ribosomal protein genes comprise a large class of genes that are repressed upon depletion of glucose. Each of the genes profiled here contains one or more RAP1-binding motifs upstream of its promoter. RAP1 is a transcriptional regulator of most ribosomal proteins.

- tion, the bound DNA was denatured by a 2-min incubation in distilled water at -95°C . The slides were then transferred into a bath of 100% ethanol at room temperature, rinsed, and then spun dry in a clinical centrifuge. Slides were stored in a closed box at room temperature until used.
10. YPD medium (8 liters), in a 10-liter fermentation vessel, was inoculated with 2 ml of a fresh overnight culture of yeast strain DBY7286 (MATa, ura3, GAL2). The fermentor was maintained at 30°C with constant agitation and aeration. The glucose content of the media was measured with a UV test kit (Boehringer Mannheim, catalog number 716251). Cell density was measured by OD at 600-nm wavelength. Aliquots of culture were rapidly withdrawn from the fermentation vessel by peristaltic pump, spun down at room temperature, and then flash frozen with liquid nitrogen. Frozen cells were stored at -80°C .
 11. Cy3-dUTP or Cy5-dUTP (Amersham) was incorporated during reverse transcription of 1.25 μg of polyadenylated [poly(A)⁺] RNA, primed by a dT(16) oligomer. This mixture was heated to 70°C for 10 min, and then transferred to ice. A premixed solution, consisting of 200 U Superscript II (Gibco), buffer, deoxyribonucleoside triphosphates, and fluorescent nucleotides, was added to the RNA. Nucleotides were used at these final concentrations: 500 μM for dATP, dCTP, and dGTP and 200 μM for dTTP. Cy3-dUTP and Cy5-dUTP were used at a final concentration of 100 μM . The reaction was then incubated at 42°C for 2 hours. Unincorporated fluorescent nucleotides were removed by first diluting the reaction mixture with 470 μl of 10 mM Tris-HCl (pH 8.0)/1 mM EDTA and then subsequently concentrating the mix to $\sim 5 \mu\text{l}$, using Centricon-30 microconcentrators (Amicon).
 12. Purified, labeled cDNA was resuspended in 11 μl of $3.5\times$ SSC containing 10 μg poly(dA) and 0.3 μl of 10% SDS. Before hybridization, the solution was boiled for 2 min and then allowed to cool to room temperature. The solution was applied to the microarray under a cover slip, and the slide was placed in a custom hybridization chamber which was subsequently incubated for ~ 8 to 12 hours in a water bath at 62°C . Before scanning, slides were washed in $2\times$ SSC, 0.2% SDS for 5 min, and then $0.05\times$ SSC for 1 min. Slides were dried before scanning by centrifugation at 500 rpm in a Beckman CS-6R centrifuge.
 13. The complete data set is available on the Internet at cmgm.stanford.edu/pbrown/explore/index.html
 14. For 95% of all the genes analyzed, the mRNA levels measured in cells harvested at the first and second interval after inoculation differed by a factor of less than 1.5. The correlation coefficient for the comparison between mRNA levels measured for each gene in these two different mRNA samples was 0.98. When duplicate mRNA preparations from the same cell sample were compared in the same way, the correlation coefficient between the expression levels measured for the two samples by comparative hybridization was 0.99.
 15. The numbers and identities of known and putative genes, and their homologies to other genes, were gathered from the following public databases: *Saccharomyces* Genome Database (genome-www.stanford.edu), Yeast Protein Database (quest7.proteome.com), and Munich Information Centre for Protein Sequences (speedy.mips.biochem.mpg.de/mips/yeast/index.htm).
 16. A. Scholer and H. J. Schuller, *Mol. Cell. Biol.* 14, 3613 (1994).
 17. S. Kratzer and H. J. Schuller, *Gene* 161, 75 (1995).
 18. R. J. Haselbeck and H. L. McAister, *J. Biol. Chem.* 268, 12116 (1993).
 19. M. Fernandez, E. Fernandez, R. Rodicio, *Mol. Gen. Genet.* 242, 727 (1994).
 20. A. Hartig et al., *Nucleic Acids Res.* 20, 5677 (1992).
 21. P. M. Martinez et al., *EMBO J.* 15, 2227 (1996).
 22. J. C. Varela, U. M. Praekelt, P. A. Meacock, R. J. Planta, W. H. Mager, *Mol. Cell. Biol.* 15, 6232 (1995).
 23. H. Ruis and C. Schuller, *Bioessays* 17, 959 (1995).
 24. J. L. Parrou, M. A. Teste, J. Francois, *Microbiology* 143, 1891 (1997).
 25. This expression profile was defined as having an induction of greater than 10-fold at 18.5 hours and less than 11-fold at 20.5 hours.
 26. S. L. Forsburg and L. Guarente, *Genes Dev.* 3, 1166 (1989).
 27. J. T. Olesen and L. Guarente, *ibid.* 4, 1714 (1990).
 28. M. Rosenkrantz, C. S. Kell, E. A. Pennell, L. J. Devenish, *Mol. Microbiol.* 13, 119 (1994).
 29. Single-letter abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr. The nucleotide codes are as follows: B-C, G, or T; N-G, A, T, or C; R-A or G; and Y-C or T.
 30. C. Fondrat and A. Kalogeropoulos, *Comput. Appl. Biosci.* 12, 363 (1996).
 31. D. Shore, *Trends Genet.* 10, 408 (1994).
 32. R. J. Planta and H. A. Raue, *ibid.* 4, 64 (1988).
 33. The degenerate consensus sequence VYCYRNNC-MNH was used to search for potential RAP1-binding sites. The exact consensus, as defined by (30), is WACAYCCRTACATYW, with up to three differences allowed.
 34. S. F. Neuman, S. Bhattacharya, J. R. Broach, *Mol. Cell. Biol.* 15, 3187 (1995).
 35. P. Lesage, X. Yang, M. Carlson, *ibid.* 16, 1921 (1996).
 36. For example, we observed large inductions of the genes coding for *PCK1*, *FBP1* [Z. Yin et al., *Mol. Microbiol.* 20, 751 (1996)], the central glyoxylate cycle gene *ICL1* [A. Scholer and H. J. Schuller, *Curr. Genet.* 23, 375 (1993)], and the "aerobic" isoform of acetyl-CoA synthase, *ACS1* [M. A. van den Berg et al., *J. Biol. Chem.* 271, 28953 (1996)], with concomitant down-regulation of the glycolytic-specific genes *PFK1* and *PFK2* [P. A. Moore et al., *Mol. Cell. Biol.* 11, 5330 (1991)]. Other genes not directly involved in carbon metabolism but known to be induced upon nutrient limitation include genes encoding cytosolic catalase *TCT1* [P. H. Bissinger et al., *ibid.* 9, 1309 (1989)] and several genes encoding small heat-shock proteins, such as *HSP12*, *HSP26*, and *HSP42* [I. Farkas et al., *J. Biol. Chem.* 266, 15602 (1991); U. M. Praekelt and P. A. Meacock, *Mol. Gen. Genet.* 223, 97 (1990); D. Wotton et al., *J. Biol. Chem.* 271, 2717 (1996)].
 37. The levels of induction we measured for genes that were expressed at very low levels in the uninduced state (notably, *FBP1* and *PCK1*) were generally lower than those previously reported. This discrepancy was likely due to the conservative background subtraction method we used, which generally resulted in overestimation of very low expression levels (46).
 38. Cross-hybridization of highly related sequences can also occasionally obscure changes in gene expression, an important concern where members of gene families are functionally specialized and differentially regulated. The major alcohol dehydrogenase genes, *ADH1* and *ADH2*, share 88% nucleotide identity. Reciprocal regulation of these genes is an important feature of the diauxic shift, but was not observed in this experiment, presumably because of cross-hybridization of the fluorescent cDNAs representing these two genes. Nevertheless, we were able to detect differential expression of closely related isoforms of other enzymes, such as *HXK1/HXK2* (77% identical) [P. Herrero et al., *Yeast* 11, 137 (1995)], *MLS1/DAL7* (73% identical) (20), and *PGM1/PGM2* (72% identical) [D. Oh, J. E. Hopper, *Mol. Cell. Biol.* 10, 1415 (1990)], in accord with previous studies. Use in the microarray of deliberately selected DNA sequences corresponding to the most divergent segments of homologous genes, in lieu of the complete gene sequences, should relieve this problem in many cases.
 39. F. E. Williams, U. Varanasi, R. J. Trumbly, *Mol. Cell. Biol.* 11, 3307 (1991).
 40. D. Tzamaras and K. Struhl, *Nature* 369, 758 (1994).
 41. Differences in mRNA levels between the *tup1 Δ* and wild-type strain were measured in two independent experiments. The correlation coefficient between the complete sets of expression ratios measured in these duplicate experiments was 0.83. The concordance between the sets of genes that appeared to be induced was very high between the two experiments. When only the 355 genes that showed at least a twofold increase in mRNA in the *tup1 Δ* strain in either of the duplicate experiments were compared, the correlation coefficient was 0.82.
 42. The *tup1 Δ* mutation consists of an insertion of the *LEU2* coding sequence, including a stop codon, between the ATG of *TUP1* and an Eco RI site 124 base pairs before the stop codon of the *TUP1* gene.
 43. L. R. Kowalski, K. Kondo, M. Inouye, *Mol. Microbiol.* 15, 341 (1995).
 44. M. Viswanathan, G. Muthukumar, Y. S. Cong, J. Lenard, *Gene* 148, 149 (1994).
 45. D. Hirata, K. Yano, T. Miyakawa, *Mol. Gen. Genet.* 242, 250 (1994).
 46. A. Gutierrez, L. Caramelo, A. Prieto, M. J. Martinez, A. T. Martinez, *Appl. Environ. Microbiol.* 60, 1783 (1994).
 47. A. Muheim et al., *Eur. J. Biochem.* 195, 369 (1991).
 48. J. A. Wemmie, M. S. Szczypka, D. J. Thiele, W. S. Moye-Rowley, *J. Biol. Chem.* 269, 32592 (1994).
 49. Microarrays were scanned using a custom-built scanning laser microscope built by S. Smith with software written by N. Ziv. Details concerning scanner design and construction are available at cmgm.stanford.edu/pbrown. Images were scanned at a resolution of 20 μm per pixel. A separate scan, using the appropriate excitation line, was done for each of the two fluorophores used. During the scanning process, the ratio between the signals in the two channels was calculated for several array elements containing total genomic DNA. To normalize the two channels with respect to overall intensity, we then adjusted photomultiplier and laser power settings such that the signal ratio at these elements was as close to 1.0 as possible. The combined images were analyzed with custom-written software. A bounding box, fitted to the size of the DNA spots in each quadrant, was placed over each array element. The average fluorescent intensity was calculated by summing the intensities of each pixel present in a bounding box, and then dividing by the total number of pixels. Local area background was calculated for each array element by determining the average fluorescent intensity for the lower 20% of pixel intensities. Although this method tends to underestimate the background, causing an underestimation of extreme ratios, it produces a very consistent and noise-tolerant approximation. Although the analog-to-digital board used for data collection possesses a wide dynamic range (12 bits), several signals were saturated (greater than the maximum signal intensity allowed) at the chosen settings. Therefore, extreme ratios at bright elements are generally underestimated. A signal was deemed significant if the average intensity after background subtraction was at least 2.5-fold higher than the standard deviation in the background measurements for all elements on the array.
 50. In addition to the 17 genes shown in Table 1, three additional genes were induced by an average of more than threefold in the duplicate experiments, but in one of the two experiments, the induction was less than twofold (range 1.6- to 1.9-fold).
 51. We thank H. Bennett, P. Spellman, J. Ravetto, M. Eisen, R. Pillai, B. Dunn, T. Ferea, and other members of the Brown lab for their assistance and helpful advice. We also thank S. Friend, D. Botstein, S. Smith, J. Hudson, and D. Dolginow for advice, support, and encouragement; K. Struhl and S. Chatterjee for the *Tup1* deletion strain; L. Fernandes for helpful advice on *Yap1*; and S. Klapholz and the reviewers for many helpful comments on the manuscript. Supported by a grant from the National Human Genome Research Institute (NHGRI) (HG00450), and by the Howard Hughes Medical Institute (HHMI). J.D.R. was supported by the HHMI and the NHGRI. V.R. was supported in part by an Institutional Training Grant in Genome Science (T32 HG00044) from the NHGRI. P.O.B. is an associate investigator of the HHMI.

5 September 1997; accepted 22 September 1997

Drug target validation and identification of secondary drug target effects using DNA microarrays

MATTHEW J. MARTON¹, JOSEPH L. DERISI², HOLLY A. BENNETT¹, VISHWANATH R. IYER²,
MICHAEL R. MEYER¹, CHRISTOPHER J. ROBERTS¹, ROLAND STOUGHTON¹, JULIA BURCHARD¹,
DAVID SLADE¹, HONGYUE DAI¹, DOUGLAS E. BASSETT, JR.¹, LELAND H. HARTWELL³,
PATRICK O. BROWN² & STEPHEN H. FRIEND¹

¹Rosetta Inpharmatics, 12040 115th Avenue NE, Kirkland, Washington 98034, USA

²Department of Biochemistry, Stanford University School of Medicine, Howard Hughes Medical Institute
Stanford, California 94305-5428, USA

³Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue N., Seattle, Washington 98109, USA

Correspondence should be addressed to S.H.F.; email: sfriend@rosetta.org

We describe here a method for drug target validation and identification of secondary drug target effects based on genome-wide gene expression patterns. The method is demonstrated by several experiments, including treatment of yeast mutant strains defective in calcineurin, immunophilins or other genes with the immunosuppressants cyclosporin A or FK506. Presence or absence of the characteristic drug 'signature' pattern of altered gene expression in drug-treated cells with a mutation in the gene encoding a putative target established whether that target was required to generate the drug signature. Drug dependent effects were seen in 'targetless' cells, showing that FK506 affects additional pathways independent of calcineurin and the immunophilins. The described method permits the direct confirmation of drug targets and recognition of drug-dependent changes in gene expression that are modulated through pathways distinct from the drug's intended target. Such a method may prove useful in improving the efficiency of drug development programs.

Good drugs are potent and specific; that is, they must have strong effects on a specific biological pathway and minimal effects on all other pathways. Confirmation that a compound inhibits the intended target (drug target validation) and the identification of undesirable secondary effects are among the main challenges in developing new drugs. Comprehensive methods that enable researchers to determine which genes or activities are affected by a given drug might improve the efficiency of the drug discovery process by quickly identifying potential protein targets, or by accelerating the identification of compounds likely to be toxic. DNA microarray technology, which permits simultaneous measurement of the expression levels of thousands of genes, provides a comprehensive framework to determine how a compound affects cellular metabolism and regulation on a genomic scale¹⁻¹¹. DNA microarrays that contain essentially every open reading frame (ORF) in the *Saccharomyces cerevisiae* genome have already been used successfully to explore the changes in gene expression that accompany large changes in cellular metabolism or cell cycle progression⁷⁻¹⁰.

In the modern drug discovery paradigm, which typically begins with the selection of a single molecular target, the ideal inhibitory drug is one that inhibits a single gene product so completely and so specifically that it is as if the gene product were absent. Treating cells with such a drug should induce changes in gene expression very similar to those resulting from deleting the gene encoding the drug's target. Here we have compared the genome-wide effects on gene expression that result from deletions of various genes in the budding yeast *S. cerevisiae* to the effects on gene expression that result from treatment

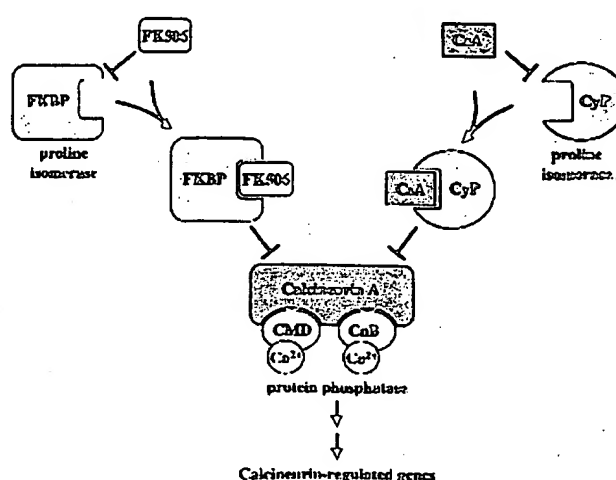
with known inhibitors of those gene products. Using the calcineurin signaling pathway as a model system, we tested an approach that permits identification of genes that encode proteins specifically involved in pathways affected by a drug. The FK506 characteristic pattern, or 'signature', of altered gene expression was not observed in mutant cells lacking proteins inhibited by FK506 (for example, a calcineurin or FK506-binding-protein mutant strain), but was observed in mutants deleted for genes in pathways unrelated to FK506 action (for example, a cyclophilin mutant strain). Conversely, the cyclosporin A (CsA) signature was not observed in CsA-treated calcineurin or cyclophilin mutant strains, but was seen in an FK506-binding-protein mutant strain treated with CsA. The method also demonstrates that FK506, a clinically used immunosuppressant, has 'off-target' effects that are independent of its binding to immunophilins. Thus, the approach we describe may provide a way to identify the pathways altered by a drug and to detect drug effects mediated through unintended targets.

Null mutants phenocopy drug-treated cells on a genomic scale

To test whether a null mutation in a drug target serves as a model of an ideal inhibitory drug, we examined the effects on gene expression associated with pharmacological or genetic inhibition of calcineurin function. Calcineurin is a highly conserved calcium- and calmodulin-activated serine/threonine protein phosphatase implicated in diverse processes dependent on calcium signaling¹²⁻¹³. In budding yeast, calcineurin is required for intracellular ion homeostasis¹⁴, for adaptation to prolonged mating pheromone treatment¹⁵ and in the regulation of

ARTICLES

Fig. 1 Model of antagonism of the calcineurin signaling pathway mediated by FK506 and cyclosporin A (CsA). Calcineurin activity is composed of a catalytic subunit (calcineurin A, encoded in yeast by the *CNA1* and *CNA2* genes), and calcium-binding regulatory subunits calmodulin (CMD) and calcineurin B (CnB). After entering cells, FK506 and CsA specifically bind and inhibit the peptidyl-proline isomerase activity of their respective immunophilins, FK506 binding proteins (FKBP) and cyclophilins (Cyp). The most abundant immunophilins in yeast (Fpr1 and Cph1) are thought to mediate calcineurin inhibition. Drug-immunophilin complexes bind and inhibit the calcium- and calmodulin-stimulated phosphatase calcineurin. Among the substrates of calcineurin are transcriptional activators that act to modulate gene expression.



the onset of mitosis¹⁶. In mammals, calcineurin has been implicated in T-cell activation¹², in apoptosis¹⁷, in cardiac hypertrophy¹⁸ and in the transition from short-term to long-term memory¹⁹. In both organisms, calcineurin activity is inhibited by FK506 and CsA, immunosuppressant drugs whose effects on calcineurin are mediated through families of intracellular receptor proteins called immunophilins^{12,20} (Fig. 1). To assess the effects of pharmacologic inhibition of calcineurin, wild-type *S. cerevisiae* was grown to early logarithmic phase in the presence or absence of FK506 or CsA. Isogenic cells, from which the genes encoding the catalytic subunits of calcineurin (*CNA1* and *CNA2*) had been deleted²¹ (referred to as the *cna* or calcineurin mutant), were grown in parallel, in the absence of the drug. Fluorescently-labeled cDNA was prepared by reverse transcription of polyA⁺ RNA in the presence of Cy3- or Cy5-deoxynucleotide triphosphates and then hybridized to a microarray containing more than 6,000 DNA probes representing 97% of the known or predicted ORFs in the yeast genome. Simultaneous hybridization of Cy5-labeled cDNA from mock-treated cells and Cy3-labeled cDNA from cells treated with 1 μ g/ml FK506 allowed the effect of drug treatment on mRNA levels of each ORF to be determined (Fig. 2a and b and data not shown). Similarly, effects of the calcineurin mutations on the mRNA levels of each gene were assessed by simultaneous hybridization of Cy5-labeled cDNA from wild-type cells and Cy3-labeled cDNA from the calcineurin mutant strain (Fig. 2c). For each comparison of this kind, reported expression ratios are the average of at least two hybridizations in which the Cy3 and Cy5 fluors were reversed to remove biases that may be introduced by gene-specific differences in incorporation of the two fluors (data not shown).

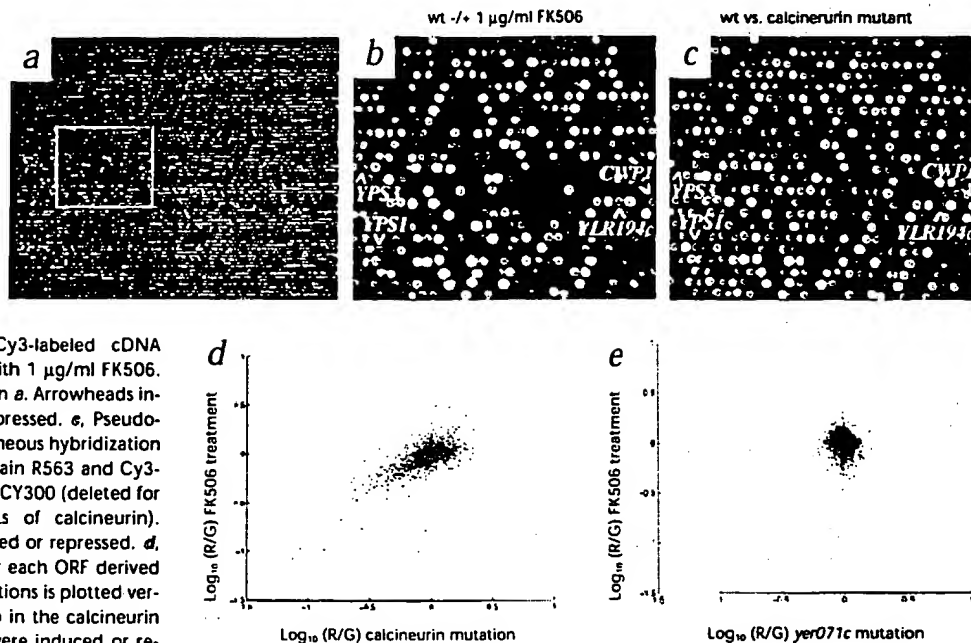
Treatment with FK506 in these growth conditions resulted in a signature pattern of altered gene expression in which mRNA levels of 36 ORFs changed by more than twofold (<http://www.rosetta.org>). A very similar pattern of altered gene expression was observed when the calcineurin mutant strain was compared to wild-type cells. Comparison of the changes in mRNA expression of each gene resulting from treatment of wild-type cells with FK506 with mRNA expression changes resulting from deletion of the calcineurin genes showed the considerable similarity of the global transcript alterations in response to the two perturbations (Fig. 2b-d). Quantification of this similarity using the correlation coefficient (ρ) showed large correlations between the FK506 treatment signature and the calcineurin deletion signature ($\rho = 0.75 \pm 0.03$), as well as the CsA treatment signature ($\rho = 0.94 \pm 0.02$), but not with a randomly selected deletion mutant strain (deleted for the *YER071C* gene; $\rho = -0.07 \pm 0.04$; Fig. 2e). The FK506 treatment signature was also compared with those of more than 40 other deletion mutant strains or drug-treatments thought to affect

unrelated pathways, and none had statistically significant correlations. These data establish that genetic disruption of calcineurin function provides a close and specific phenocopy of treatment with FK506 or CsA.

To avoid generalizing from a single example, we also compared the effects of treatment of wild-type cells with 3-aminotriazole (3-AT) with the effects of deletion of the *HIS3* gene. *HIS3* encodes imidazoleglycerol phosphate dehydratase, which catalyzes the seventh step of the histidine biosynthetic pathway in yeast²²; 3-AT is a competitive inhibitor of this enzyme that triggers a large transcriptional amino-acid starvation response²³. Microarray analysis of wild-type and isogenic *his3*-deficient strains demonstrated the expected large genome-wide transcriptional responses (involving more than 1,000 ORFs) resulting from treatment with 3-AT (Fig. 3a) or from *HIS3* deletion (Fig. 3c). Quantitative comparison of the 3-AT treatment signature and the *his3* mutant signature showed a high level of correlation ($\rho = 0.76 \pm 0.02$) that even extended to genes that experienced small changes in expression level (Fig. 3b). As a negative control, the correlations between the 3-AT treatment signature or the *his3* mutant signature and the calcineurin mutant strain were not statistically significant ($\rho = 0.09 \pm 0.06$ and -0.01 ± 0.04 , respectively). That both the calcineurin/FK506 and the *his3*/3-AT comparisons were highly correlated indicates that in many cases the expression profile resulting from a gene deletion closely resembles the expression profile of wild-type cells treated with an inhibitor of that gene's product.

'Decoder' strategy: Drug target validation with deletion mutants
Because pharmacological inhibition of different targets might give similar or identical expression profiles, simple comparison of drug signatures to mutant signatures is unlikely to unambiguously identify a drug's target. To overcome this limitation, an additional 'decoder' step is used. We first compare the expression profile of wild-type drug-treated cells to the expression profiles from a panel of genetic mutant strains, using a correlation coefficient metric. Mutant strains whose expression profile is similar to that of drug-treated wild-type cells are selected and subjected to drug treatment, generating the drug signature in the mutant strain (that is, the mutant drug signature). If the mutated gene encodes a protein involved in a pathway affected by the drug, we expect the drug signature in mutant cells to be different (or absent, for an ideal drug) from the drug signature seen in wild-type cells.

Fig. 2 Expression profiles from FK506-treated wild-type (wt) cells and a calcineurin-disruption mutant strain share a genome-wide correlation. DNA microarray analysis showing changes in gene expression resulting from FK506 treatment (a and b) or from genetic disruption of genes encoding calcineurin (c). a, Pseudocolor image of the results of simultaneous hybridization of Cy5-labeled cDNA (red) from mock-treated strain R563 and Cy3-labeled cDNA (green) from strain R563 treated with 1 μ g/ml FK506. b, Enlarged view of the boxed area in a. Arrowheads indicate specific ORFs induced or repressed. c, Pseudocolor image of the results of simultaneous hybridization of Cy5-labeled cDNA (red) from strain R563 and Cy3-labeled cDNA (green) from strain MCY300 (deleted for the *CNA1*, *CNA2* catalytic subunits of calcineurin). Arrows indicate specific ORFs induced or repressed. d, The \log_{10} of the expression ratio for each ORF derived from the FK506 treatment hybridizations is plotted versus the \log_{10} of the expression ratio in the calcineurin mutant hybridizations. ORFs that were induced or repressed in both experiments are shown as green and red dots, respectively. e, The \log_{10} of the expression ratio for each ORF derived from the FK506 treatment hybridizations is plotted versus the \log_{10}



of the expression ratio in the *yer071c* mutant hybridizations. No ORFs were induced or repressed in both experiments.

To illustrate this, we treated the *his3* mutant strain with 3-AT. The signature pattern of altered gene expression resulting from treatment of the mutant strain with 3-AT was much less complex than that of the 3-AT signature in wild-type cells (Fig. 4). This is seen simply by examining plots of mean intensity of the hybridization signal (which approximately reflects level of expression) versus the expression ratio for each ORF (Fig. 4). Genes that were expressed at higher or lower levels in 3-AT treated cells or in *his3* mutant cells are shown as red and green dots, respectively. We analyzed the 3-AT signature in wild-type (Fig. 4a) and *his3* mutant cells (Fig. 4c), as well as the *his3* mutant strain signature (Fig. 4b). Whereas histidine limitation induced by 3-AT induced more than 1,000 transcription-level changes in the wild-type strain, few or no transcript level changes were induced by treatment of the *his3*-deletion strain with 3-AT. This indicates that with the growth conditions used, essentially all of the effects of 3-AT depend on or are mediated through the *HIS3* gene product.

Applying this approach to the calcineurin signaling pathway showed the specificity of the method. The calcineurin mutant strain and strains with deletions in the genes encoding the most abundant immunophilins in yeast¹² (*CPH1* and *FPR1*) were treated with either FK506 or CsA to determine the profiles

of altered gene expression resulting from drug treatment of the mutant cells (that is, mutant +/- drug). We compared the drug signatures in the mutants to the wild-type drug signature using the correlation coefficient metric (Table 1). Although the signature generated by treatment of wild-type cells with FK506 was highly correlated to the calcineurin mutant strain signature ($\rho = 0.75 \pm 0.03$), it bore no similarity to the profile after treatment of the calcineurin mutant strain with FK506 ($\rho = -0.01 \pm 0.07$). This indicates that FK506 was unable to elicit its normal transcriptional response in the calcineurin mutant strain. Likewise, treatment of the *fpr1* mutant strain with FK506 elicited an expression profile that was not correlated to the FK506 signature in the wild-type strain ($\rho = -0.23 \pm 0.07$), indicating that the *FPR1* gene product is likely to be involved in the pathway affected by FK506. The same was true for the *cna fpr1* mutant strain. In contrast, treatment of the *cph1* mutant strain with FK506 generated an expression profile highly correlated with the wild-type FK506 expression profile ($\rho = 0.79 \pm 0.03$), indicating the *cph1* mutation did not block the mode of action of FK506 and thus is not directly involved in the pathway affected by FK506. We tabulated the change in expression in response to FK506 in different mutant strains for all ORFs with expression ratios greater than 1.8 in FK506-treated cells or in the calcineurin mutant strain (Fig. 5a). The calcineurin mutant strain signature and the FK506 responses in wild-type and the *cph1* mutant strain are similar, and there are no transcript-level changes (seen in black) for treatment of the calcineurin, *fpr1* and *cna fpr1* mutant strains with FK506 (Fig. 5a).

Similar experiments and analyses with CsA provided further validation of this approach. The expression profile elicited by treatment of wild-type cells with CsA was highly corre-

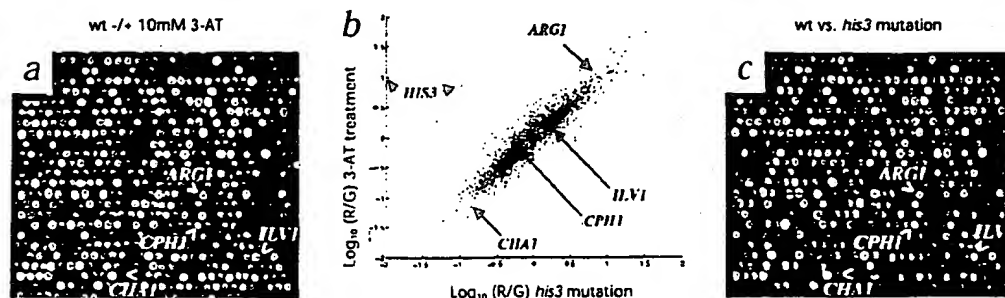
Table 1 Signature correlation of expression ratios as a result of FK506 treatment in various mutant strains

	wild-type +/-FK506	<i>cna</i> +/-FK506	<i>fpr1</i> +/-FK506	<i>cna fpr1</i> +/-FK506	<i>cph1</i> +/-FK506
wild-type +/- FK506	0.93 \pm 0.04	-0.01 \pm 0.07	-0.23 \pm 0.07	0.12 \pm 0.07	0.79 \pm 0.03

Signature correlation shows the absence of the FK506 signature specifically in the calcineurin (*cna*) and *fpr1* (major FK506 binding protein) deletion mutants. *cna* represents the mutant with deletions of the catalytic subunits of calcineurin, *CNA1* and *CNA2*. The correlation coefficient reported in the first column represents the correlation between two pairs of hybridizations from independent wild-type +/- FK506 experiments.

ARTICLES

Fig. 3 Expression profiles from a *his3* mutant strain and wild-type (wt) cells treated with 3-AT share a genome-wide correlation. DNA microarray analysis showing changes in gene expression resulting from 3-AT treatment (a) or from genetic disruption of the *HIS3* gene (c). a, Pseudo-color image of the results of simultaneous hybridization of Cy5-labeled cDNA (red) from mock-treated wild-type strain R491 and Cy3-labeled cDNA (green) from strain R491 treated with 10 mM 3-AT. b, Plot of the \log_{10} of the expression ratio for each ORF derived from the 3-AT treatment hybridizations is plotted versus the \log_{10} of the expression ratio in the *his3* mutant hybridizations. ORFs that were induced or repressed in both experiments are shown as green and red dots, respectively. The correlation of expression ratios applies not only to genes with large expression ratios (for example, *CHA1* and *ARG1*), but also extends to genes with expression ratios less than 2 (for example, *ILV1* and *CPH1*). *ILV1* is induced 1.9-fold and 1.5-fold, and *CPH1* is downregulated 1.9-fold



and 1.7-fold, in cells treated with 3-AT and *his3* mutant cells, respectively. Two ORFs do not fall on the line $x = y$. The leftmost point is the *HIS3* data point, which is induced by 3-AT treatment but which is not absent from the *his3* mutant strain. The other point is *YOR203w*. Both data points are labeled *HIS3* because hybridization to *YOR203w* is most likely due to *HIS3* mRNA, as *YOR203w* overlaps the *HIS3* open reading frame. c, Pseudo-color image of the results of simultaneous hybridization of Cy5-labeled cDNA (red) from wild-type strain R491 and Cy3-labeled cDNA (green) from strain R1226, deleted for the *HIS3* gene. Arrowheads indicate specific ORFs induced or repressed.

lated to the profile elicited by mutation of the calcineurin genes ($p = 0.71 \pm 0.04$), but did not correlate with the expression profile resulting from treatment of the calcineurin mutant strain with CsA ($p = -0.05 \pm 0.07$; Table 2), indicating that the genetic deletion of calcineurin interfered with the ability of CsA to elicit its normal transcriptional response. Likewise, the CsA signature was essentially absent in CsA-treated *cph1* mutant cells, and the expression profile of CsA-treated *cph1* mutant cells correlated poorly to that of CsA-treated wild-type cells ($p = 0.18 \pm 0.07$). Thus, the *CPH1* gene product was required for the CsA response seen in wild-type cells. Conversely, treatment of *fpr1* mutant cells with CsA resulted in an expression pattern very similar to the profile of CsA-treated wild-type cells ($p = 0.77 \pm 0.03$), indicating that *FPR1* was not necessary for the CsA-mediated effects. Analysis of individual ORFs affected by CsA and their expression ratios over the entire set of experiments confirmed that *CPH1* and the genes encoding calcineurin, but not

FPR1, are necessary for the wild-type CsA response (Fig. 5b). The observation that the profiles resulting from FK506 or CsA drug treatment are similar to that of the calcineurin deletion mutant strain might allow the prediction that calcineurin was involved in the pathway affected by these drugs. But because the expression profile of the *fpr1* mutant strain did not bear a strong similarity to the wild-type drug expression profile for FK506, it is obvious that the drug treatment of the mutant strains was necessary to identify *Fpr1*, but not *Cph1*, as a potential FK506 drug target. In the same way, the 'decoder' strategy was necessary to identify *Cph1*, but not *Fpr1*, as a potential drug target for CsA.

'Decoder' approach can identify secondary drug effects

For a drug that has a single biochemical target, the strategy outlined above may be useful in target validation. In many cases, however, a compound may affect multiple pathways and elicit a very complex signature. 'Decoding' such a complex signature

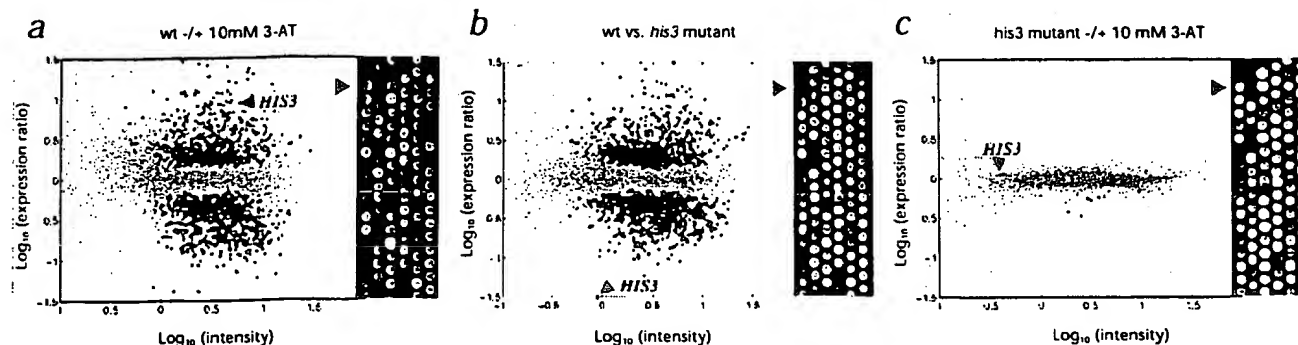


Fig. 4 Treatment of the *his3* mutant strain with 3-AT shows nearly complete loss of 3-AT signature. a, Plot of the \log_{10} of the mean intensity of hybridization for each ORF versus the \log_{10} of its expression ratio for each experiment is shown next to a pseudo-color image of a representative portion of the microarray. ORFs that are induced or repressed at the 95% confidence level are shown in green and red, respectively. b, Expression profile from treatment of the wild-type (wt) strain with 3-AT. Cy5-labeled cDNA (red) from mock-treated strain R491 and Cy3-labeled cDNA (green) from strain R491 treated with 10 mM 3-AT. c, Expression profile

from the *his3* deletion strain. Cy5-labeled cDNA (red) from strain R491 and Cy3-labeled cDNA (green) from strain R1226, deleted for the *HIS3* gene. d, Expression profile of treatment of the *his3* deletion strain with 3-AT. Cy5-labeled cDNA (red) from *his3*-deleted strain R1226 and Cy5-labeled cDNA (green) from strain R1226 treated with 10 mM 3-AT. Arrowheads indicate the DNA probe and data point corresponding to the *HIS3* gene. The blue dashed line represents the threshold below which errors tend to increase rapidly because spot intensities are not sufficiently above background intensity.

Table 2 Signature correlation of expression ratios as a result of CsA treatment in various mutant strains

	wild-type +/-CsA	<i>cna</i> +/-CsA	<i>fpr1</i> +/-CsA	<i>cna cph1</i> +/-CsA	<i>cph1</i> +/-CsA
wild-type +/- CsA	0.94 ± 0.04	-0.05 ± .07	0.77 ± 0.03	-0.11 ± 0.07	0.18 ± 0.07

Signature correlation shows the absence of the CsA signature specifically in the calcineurin (*cna*) and *cph1* (cyclophilin) deletion mutants. *cna* represents the mutant with deletions of the catalytic subunits of calcineurin, *CNA1* and *CNA2*. The correlation coefficient reported in the first column represents the correlation between two pairs of hybridizations from independent wild-type +/- CsA experiments.

into the effects mediated through the intended target (the 'on-target signature') and those mediated through unintended targets (the 'off-target' signature) might be useful in evaluating a compound's specificity. Our 'decoder' strategy is based on the premise that 'off-target' signature should be insensitive to the genetic disruption of the primary target.

To determine whether the 'decoder' approach could identify an 'off-target' profile, we looked for a drug-responsive gene whose expression is insensitive to deletion of the primary target. To increase the likelihood of observing such genes, the same strains described in Tables 1 and 2 were treated with higher concentrations (50 µg/ml) of FK506. This led to a much more complex expression profile in wild-type cells, indicating that at this higher concentration, FK506 was inhibiting or activating additional targets. Several of the ORFs in this expanded FK506-induced expression profile were not affected by the calcineurin, *cph1* or *fpr1* mutations, as drug treatment of these mutant strains did not block their presence in the FK506 expression signature (Fig. 6). This indicates that FK506 was triggering changes in transcript levels of many genes through pathways independent of calcineurin, *CPH1* and *FPR1*. Many of the upregulated ORFs in the 'off-target' pathway were genes reported to be regulated by the transcriptional activator Gcn4 (ref. 24). In some strains, a reporter gene under *GCN4* control was induced in response to FK506 treatment²⁵. To determine whether *GCN4* is involved in this pathway that is independent of calcineurin, *CPH1* and *FPR1*, we analyzed the effects of treatment with high-dose FK506 on global gene expression in a strain with a *GCN4* deletion (Fig. 6). Of the 41 ORFs with calcineurin-independent expression ratios greater than 4, 32 were not induced in the *gcn4* mutant, indicating that their induction by FK506 was *GCN4*-dependent. Not all *GCN4*-regulated genes were induced by FK506. This FK506-induced subset of *GCN4*-regulated genes may be those most sensitive to subtle changes in Gcn4 levels, or perhaps other regulatory circuits prevent FK506 activation of some *GCN4*-regulated genes. Seven of the remaining nine ORFs induced by FK506 were independent of

both the calcineurin and *GCN4* pathways. The simplest explanation is that FK506 inhibits or activates additional pathways. Members of this class include *SNQ2* and *PDR5*, genes that encode drug efflux pumps with structural homology to mammalian multiple drug resistance proteins²⁶. FK506 may interact directly with Pdr5 to inhibit its function²⁷. Our results indicate that treatment with FK506 leads to fourfold-to-sixfold induction of *PDR5* mRNA levels. *YOR1*, another gene that can confer drug resistance, is also induced threefold-to-fourfold by

FK506. Thus, drug treatment of strains with mutations in the primary targets can prove useful in identifying effects mediated by secondary drug targets, including the nature and extent of newly discovered and previously unsuspected pathways affected by the drug.

We describe here a method for drug target validation and the identification of secondary drug target effects that uses DNA microarrays to survey the effects of drugs on global gene expression patterns. We established that genetic and pharmacologic inhibition of gene function can result in extremely similar changes in gene expression. We also demonstrated that one can confirm a potential drug target by treating a deletion mutant defective in the gene encoding the putative target. Drug-mediated signatures from strains with mutations in pathways or processes directly or indirectly affected by the drug bore little or

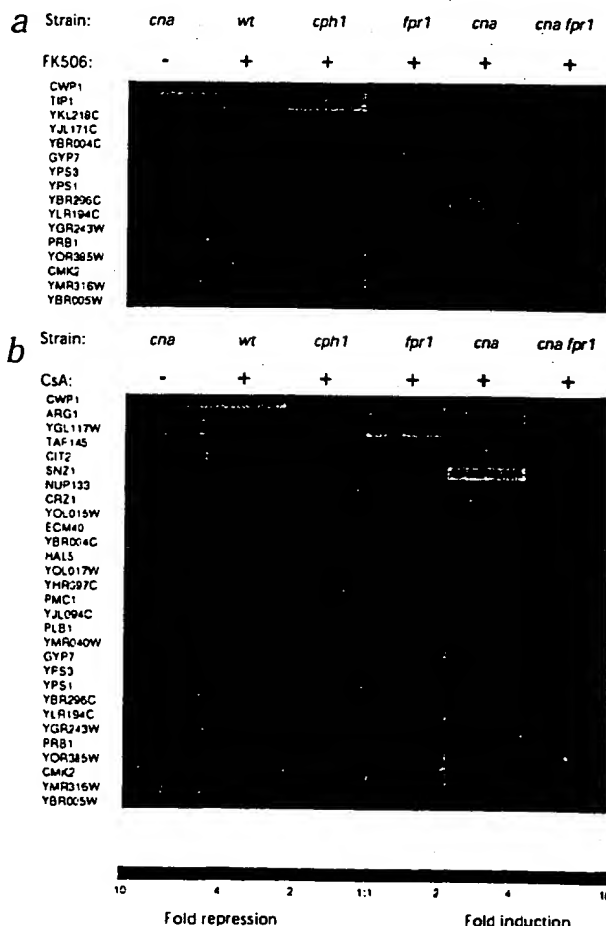


Fig. 5 Response of FK506 and CsA signature genes in strains with deletions in different genes. Genes with expression ratios greater than a factor of 1.8 in response to treatment with 1 µg/ml FK506 (a) or 50 µg/ml CsA (b) are listed (left side) and their expression ratios in the indicated strain are shown on the green (induction)-red (repression) color scale. **a**, Calcineurin (*cna*) mutant and FK506 treatment signature genes are in the first two columns. Almost all FK506 signature genes have expression ratios near unity in deletion strains involved in pathways affected by FK506 (calcineurin, *fpr1* and *cna fpr1* mutants) but not in deletion strains in unrelated pathways (*cph1*). **b**, Calcineurin (*cna*) mutant and CsA treatment signature genes are in the first two columns. Almost all CsA signature genes have expression ratios near unity in deletion strains involved in pathways affected by CsA (calcineurin, *cph1* and *cna cph1* mutants) but not in deletion strains in unrelated pathways (*fpr1*).

ARTICLES

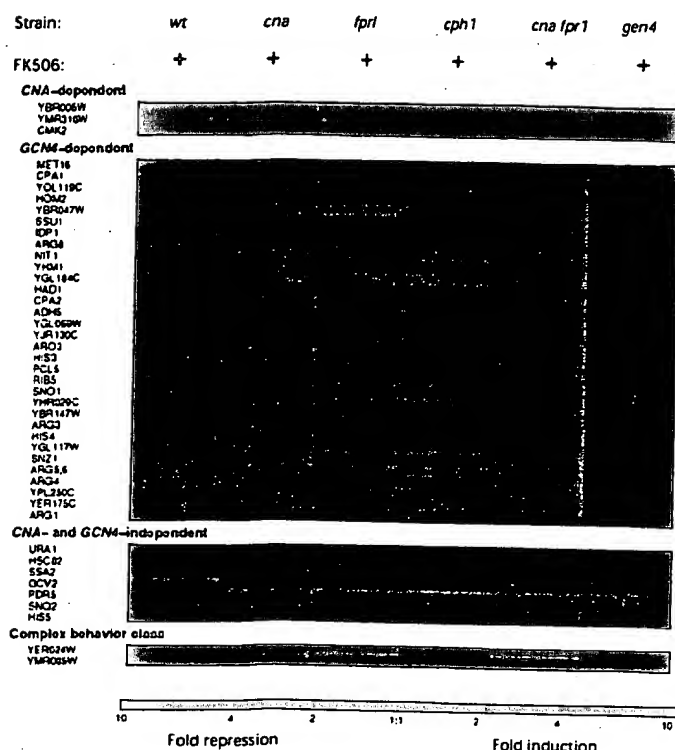


Fig. 6 Response of FK506 signature genes in strains with deletions in different genes. Genes with expression ratios greater than a factor of 4 in at least one experiment are listed and their expression ratios in the indicated strain are shown in the green (induction)–red (repression) color scale. The genes have been divided into classes corresponding to these expected behaviors: 'CNA-dependent' genes respond to FK506 (50 μ g/ml) except when either calcineurin genes or *FPR1* or both are deleted; 'GCN4-dependent' genes respond to FK506 when calcineurin genes or *FPR1* or *CPH1* are deleted; that is, their responses are not mediated by calcineurin, Cph1, or Fpr1. 'CNA- and GCN4-independent' genes respond to FK506 in all deletion strains tested. A 'complex behavior' class is provided for those genes that did not match the model of FK506 response mediated through calcineurin or Fpr1 or separately through Gcn4.

penile erection. It is possible that application of the 'decoder' to other compounds may show that they too have a potent activity against a target distinct from their intended target.

The ability to decode drug effects is dependent on the availability of functionally 'targetless' cells. In yeast, this is being achieved by systematically disrupting each yeast gene (*Saccharomyces* Deletion Consortium; http://sequence-www.stanford.edu/group/yeast_deletion_project/deletion.html). Efforts are underway to obtain expression profiles from each deletion mutant strain. Determining signatures resulting from inactivation of essential genes presents a unique problem, but it may be possible to do so by examining heterozygotes or by using a controllable promoter to reduce expression of the essential gene. Although it is already feasible to test several compounds in dozens of yeast strains, another challenge for the 'decoder' strategy will be the efficient selection of the mutants with deletions in genes most likely to encode the intended drug target. The signature correlation plots described are one metric that could be used as part of that selection process, but others need to be explored. Applying the 'decoder' to mammalian cells presents additional challenges. It is considerably more difficult to isolate functionally 'targetless' cells. Strategies involving titratable promoters, known specific inhibitors, anti-sense RNAs, ribozymes, and methods of targeting specific proteins for degradation are possible and should be tested. Another limitation is that not all cell types express the same set of genes and therefore 'off-target' effects may be different in different cell types. In addition, applying the 'decoder' to human cells will also require technical improvements that allow expression profiling from a small number of cells. Even the broader question of whether the insensitivity of 'off-target' signatures to the disruption of the main target is the exception or the rule can only be answered by the accumulation of more data. Barkai and Leibler, however, have argued in favor of robustness of biological networks, indicating that drug perturbations ('off-target' signatures) may be robust even when the system is subjected to another perturbation (such as a genetic disruption) (ref. 28). Many practical developments will be necessary if the 'decoder' concept is to be broadly applied.

Expression arrays have been used mainly as an initial screen for genes induced in a particular tissue or process of interest by focusing on genes with large expression ratios. We have found, however, that effort to refine experimental protocols and repeat experiments increases the reliability of the data and permits new applications. For example, it provides a larger set

no similarity to the wild-type drug expression profile. In contrast, drug-mediated signatures from strains with mutations in genes involved in pathways unrelated to the drug's action showed extensive similarity to the wild-type drug signature. By applying this approach to a drug that affects multiple pathways (FK506), we were able to decode a complex signature into component parts, including the identification of an 'off-target' signature that was mediated through pathways independent of calcineurin or the Fpr1 immunophilin.

Discussion

It is well-established that high-throughput biochemical screening can identify potent inhibitory compounds against a given target. The 'decoder' approach described here complements this process by evaluating the equally important property of specificity: the tendency of a compound to inhibit pathways other than that of its intended target. The ability to observe such 'off-target' effects will likely be useful in several ways. Profiling compounds with known toxicities will allow the development of a database of expression changes associated with particular toxicities. Recognition of potential toxicities in the 'off-target' signatures of otherwise promising compounds then may allow earlier identification of those likely to fail in clinical trials. Comparing the extent and peculiarities of 'off-target' signatures of promising drug candidates could provide a new way to group compounds by their effects on secondary pathways, even before those effects are understood. This may prove to be an alternative, potentially more effective, way to select compounds for animal and clinical trials. Some drugs are more effective against a related protein than against the originally intended target. Sildenafil (ViagraTM), for example, was initially developed as a phosphodiesterase inhibitor to control cardiac contractility, but was found to be highly specific for phosphodiesterase 5, an isozyme whose inhibition overcomes defects in

Table 3 Yeast strains used

Strain	Relevant genotype	Reference
YPH499	<i>Mata ura3-52 lys2-801 ade2-101 trp1-Δ63 his3-Δ200 leu2-Δ1</i>	(34)
R563	<i>Mata ura3-52 lys2-801 ade2-101 trp1-Δ63 his3-Δ200 leu2-Δ1 his3::HIS3</i>	(this study)
R558	<i>Mata ura3-52 lys2-801 ade2-101 trp1-Δ63 his3-Δ200 leu2-Δ1 fpr1::HIS3</i>	(this study)
R567	<i>Mata ura3-52 lys2-801 ade2-101 trp1-Δ63 his3-Δ200 leu2-Δ1 cph1::HIS3</i>	(this study)
MCY300	<i>Mata ura3-52 lys2-801 ade2-101 trp1-Δ63 his3-Δ200 leu2-Δ1 cna1Δ1::hisG cna2Δ1::HIS3</i>	(21)
R132	<i>Mata ura3-52 lys2-801 ade2-101 trp1-Δ63 his3-Δ200 leu2-Δ1 cna1Δ1::hisG cna2Δ1::HIS3 cph1::karf</i>	(this study)
R133	<i>Mata ura3-52 lys2-801 ade2-101 trp1-Δ63 his3-Δ200 leu2-Δ1 cna1Δ1::hisG cna2Δ1::HIS3 fpr1::karf</i>	(this study)
R559	<i>Mata ura3-52 lys2-801 ade2-101 trp1-Δ63 his3-Δ200 leu2-Δ1 his3::HIS3 gcn4::LEU2</i>	(this study)
BY4719	<i>Mata trp1-Δ63 ura3-Δ0</i>	(35)
BY4738	<i>Mata trp1-Δ63 ura3-Δ0</i>	(35)
R491	<i>Mata/α BY4719 X BY4738</i>	(this study)
BY4728	<i>Mata his3-Δ200 trp1-Δ63 ura3-Δ0</i>	(35)
BY4729	<i>Mata his3-Δ200 trp1-Δ63 ura3-Δ0</i>	(35)
R1226	<i>Mata/α BY4728 X BY4729</i>	(this study)

of genes at higher confidence levels that serve as a more unique signature for a given protein perturbation. In addition, it allows subtle signatures to be detected, when, for example, a protein is only partially inhibited. This may enable clinical monitoring of small changes in protein function in disease or toxicity states before they could otherwise be detected. Because the functions of many genes detected on transcript arrays are known, these microarrays are powerful tools that provide detailed information about a cell's physiology. For example, changes in the flux through a metabolic pathway are reflected in transcriptional changes in genes in the pathway⁷. Furthermore, it may be possible to indirectly measure protein activity levels from expression profiling data (S.F., *et al.*, unpublished data). Thus, although the eventual development of genomic methods allowing the direct measurement of all cellular protein levels will be an important achievement, transcript array technology offers an immediate and robust means of evaluating the effects of various treatments on gene expression and protein function.

Methods

Construction, growth and drug treatment of yeast strains. The strains used in this study (Table 3) were constructed by standard techniques²⁹. To construct strain R559, strain R563 was transformed to *Leu*⁻ with plasmid pM12 digested by *SalI* and *MluI* (provided by A. Hinnebusch and T. Dever). Strains R132 and R133 were constructed by transforming the bacterial kanamycin resistance cassette³⁰ flanked by genomic DNA from the *CPH1* and *FPR1* loci, respectively, and selecting for G418-resistant colonies. For experiments with FK506, cells were grown for three generations to a density of 1×10^7 cells/ml in YAPD medium (YPD plus 0.004% adenine) supplemented with 10 mM calcium chloride as described³¹. Where indicated, FK506 was added to a final concentration of 1 μg/ml 0.5 h after inoculation of the culture or to 50 μg/ml 1 h before cells were collected. CsA was used at a final concentration of 50 μg/ml. Cells were broken by standard procedures³² with the following modifications: Cell pellets were resuspended in breaking buffer (0.2 M Tris HCl pH 7.6, 0.5 M NaCl, 10 mM EDTA, 1% SDS), vortexed for 2 min on a VWR multi-tube vortexer at setting 8 in the presence of 60% glass beads (425–600 μm mesh; Sigma) and phenol:chloroform (50:50, volume/volume). After separation of the phases, the aqueous phase was re-extracted and ethanol-precipitated. Poly A⁺ RNA was isolated by two sequential chromatographic purifications over oligo dT cellulose (New England Biolabs, Beverly, Massachusetts) using established protocols³³.

For experiments using 3-AT, wild-type or *his3/his3* cells were grown to early logarithmic phase in SC medium, pelleted and resuspended in SC medium lacking histidine for 1 hr in the presence or absence of 10 mM 3-

AT, as indicated. Cells were harvested and mRNA isolated as above. FK506 was obtained from the Swedish Hospital Pharmacy (Seattle, Washington) and purified to homogeneity by ethyl acetate extraction by J. Simon (Fred Hutchinson Cancer Research Center, Seattle, Washington). CsA was obtained from Alexis Biochemicals (San Diego, California); 3-AT was from Sigma.

Preparation and hybridization of the labeled sample. Fluorescently-labeled cDNA was prepared, purified and hybridized essentially as described⁷. Cy3- or Cy5-dUTP (Amersham) was incorporated into cDNA during reverse transcription (Superscript II; Life Technologies) and purified by concentrating to less than 10 μl using Microcon-30 microconcentrators (Amicon, Houston, Texas). Paired cDNAs were resuspended in 20–26 μl hybridization solution (3 × SSC, 0.75 μg/ml polyA DNA, 0.2% SDS) and applied to the microarray under a 22 × 30-mm coverslip for 6 h at 63 °C, all according to a published method⁷.

Fabrication and scanning of microarrays. PCR products containing common 5' and 3' sequences (Research Genetics, Huntsville, Alabama) were used as templates with amino-modified forward primer and unmodified reverse primers to PCR amplify 6,065 ORFs from the *S. cerevisiae* genome. Our first-pass success rate was 94%. Amplification reactions that gave products of unexpected sizes were excluded from subsequent analysis. ORFs that could not be amplified from purchased templates were amplified from genomic DNA. DNA samples from 100-μl reactions were isopropanol-precipitated, resuspended in water, brought to a final concentration of 3 × SSC in a total volume of 15 μl, and transferred to 384-well microtiter plates (Genetix Limited, Christchurch, Dorset, England). PCR products were spotted onto 1 × 3-inch polylysine-treated glass slides by a robot built essentially according to defined specifications^{34,37} (<http://cmgm.stanford.edu/pbrown/MGuide>). After being printed, slides were processed according to published protocols⁷.

Microarrays were imaged on a prototype multi-frame CCD camera in development at Applied Precision (Issaquah, Washington). Each CCD image frame was approximately 2-mm square. Exposure times of 2 s in the Cy5 channel (white light through Chroma 618–648 nm excitation filter, Chroma 657–727 nm emission filter) and 1 s in the Cy3 channel (Chroma 535–560 nm excitation filter, Chroma 570–620 nm emission filter) were done consecutively in each frame before moving to the next, spatially contiguous frame. Color isolation between the Cy3 and Cy5 channels was about 100:1 or better. Frames were 'knitted' together in software to make the complete images. The intensity of spots (about 100 μm) were quantified from the 10-μm pixels by frame-by-frame background subtraction and intensity averaging in each channel. Dynamic range of the resulting spot intensities was typically a ratio of 1,000 between the brightest spots and the background-subtracted additive error level. Normalization between the channels was accomplished by normalizing each channel to the mean intensities of all genes. This procedure is nearly equivalent to normalization between channels using the intensity

ARTICLES

ratio of genomic DNA spots⁷, but is possibly more robust, as it is based on the intensities of several thousand spots distributed over the array.

Signature correlation coefficients and their confidence limits. Correlation coefficients between the signature ORFs of various experiments were calculated using:

$$\rho = \sum_k x_k y_k / (\sum_k x_k^2 \sum_k y_k^2)^{1/2}$$

where x_k is the \log_{10} of the expression ratio for the k^{th} gene in the x signature, and y_k is the \log_{10} of the expression ratio for the k^{th} gene in the y signature. The summation is over those genes that were either up- or down-regulated in either experiment at the 95% confidence level. These genes each had a less than 5% chance of being actually unregulated (having expression ratios departing from unity due to measurement errors alone). This confidence level was assigned based on an error model which assigns a lognormal probability distribution to each gene's expression ratio with characteristic width based on the observed scatter in its repeated measurements (repeated arrays at the same nominal experimental conditions) and on the individual array hybridization quality. This latter dependence was derived from control experiments in which both Cy3 and Cy5 samples were derived from the same RNA sample. For large numbers of repeated measurements the error reduces to the observed scatter. For a single measurement the error is based on the array quality and the spot intensity.

Random measurement errors in the x and y signatures tend to bias the correlation towards zero. In most experiments, most genes are not significantly affected but do show small random measurement errors. Selecting only the '95% confidence' genes for the correlation calculation, rather than the entire genome, reduces this bias and makes the actual biological correlations more apparent.

Correlations between a profile and itself are unity by definition. Error limits on the correlation are 95% confidence limits based on the individual measurement error bars, and assuming uncorrelated errors²³. They do not include the bias mentioned above; thus, a departure of ρ from unity does not necessarily mean that the underlying biological correlation is imperfect. However, a correlation of 0.7 ± 0.1 , for example, is very significantly different from zero. Small (magnitude of $\rho < 0.2$) but formally significant correlation in the tables and text probably are due to small systematic biases in the Cy5/Cy3 ratios that violate the assumption of independent measurement errors used to generate the 95% confidence limits. Therefore, these small correlation values should be treated as not significant. A likely source of uncorrected systematic bias is the partially corrected scanner detector nonlinearity that differently affects the Cy3 and Cy5 detection channels.

The 1 $\mu\text{g}/\text{ml}$ FK506 treatment signature was compared with more than 40 unrelated deletion mutant strain or drug signatures. These control profiles had correlation coefficients with the FK506 profile that were distributed around zero (mean $\rho = -0.03$) with a standard deviation of 0.16 (data not shown), and none had correlations greater than $\rho = 0.38$. Similarly, the calcineurin mutant strain signature correlated well with the CsA treatment signature ($\rho = 0.71 \pm 0.04$) but not with the signatures from the negative controls (mean $\rho = -0.02$ with a standard deviation of 0.18).

Quality controls. End-to-end checks on expression ratio measurement accuracy were provided by analyzing the variance in repeated hybridizations using the same mRNA labeled with both Cy3 and Cy5, and also using Cy3 and Cy5 mRNA samples isolated from independent cultures of the same nominal strain and conditions. Biases undetected with this procedure, such as gene-specific biases presumably due to differential incorporation of Cy3- and Cy5-dUTP into cDNA, were minimized by doing hybridizations in fluor-reversed pairs, in which the Cy3/Cy5 labeling of the biological conditions was reversed in one experiment with respect to the other. The expression ratio for each gene is then the ratio of ratios between the two experiments in the pair. Other biases are removed by algorithmic numerical de-trending. The magnitude of these biases in the absence of de-trending and fluor reversal is typically about 30% in the ratio, but may be as high as twofold for some ORFs.

Expression ratios are based on mean intensities over each spot. Some

smaller spots have fewer image pixels in the average. This does not degrade accuracy noticeably until the number of pixels falls below ten, in which case the spot is rejected from the data set. 'Wander' of spot positions with respect to the nominal grid is adaptively tracked in array subregions by the image processing software. Unequal spot 'wander' within a subregion greater than half-a-spot spacing is a difficulty for the automated quantitating algorithms; in this case, the spot is rejected from analysis based on human inspection of the 'wander'. Any spots partially overlapping are excluded from the data set. Less than 1% of spots typically are rejected for these reasons.

Acknowledgments

The authors thank all the members of Rosetta for their contributions to this work. We thank P. Linsley, D. Shoemaker and A. Murray for critical reading of the manuscript, and M. Cyert for providing yeast strains. Work done at Stanford was supported in part by the Howard Hughes Medical Institute, and by a grant to P.O.B from the NHGRI. P.O.B is an assistant investigator of the Howard Hughes Medical Institute.

RECEIVED 13 AUGUST; ACCEPTED 2 OCTOBER 1998

- Schena, M., Shalon, D., Davis, R.W. & Brown, P.O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270, 467-470 (1995).
- Schena, M. *et al.* Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proc. Natl. Acad. Sci. USA* 93, 10614-10619 (1996).
- Shalon, D., Smith, S.J. & Brown, P.O. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res.* 6, 639-645 (1996).
- Lockhart, D.J. *et al.* Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnol.* 14, 1675-1680 (1996).
- DeRisi, J. *et al.* Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nature Genet.* 14, 457-460 (1996).
- Heller, R.A. *et al.* Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. *Proc. Natl. Acad. Sci. USA* 94, 2150-2155 (1997).
- DeRisi, J.L., Iyer, V.R. & Brown, P.O. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278, 680-686 (1997).
- Lashkari, D.A. *et al.* Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proc. Natl. Acad. Sci. USA* 94, 13057-13062 (1997).
- Wodicka, L., Dong, H., Mittman, M., Ho, M.-H. & Lockhart, D.J. Genome-wide expression monitoring in *Saccharomyces cerevisiae*. *Nature Biotechnol.* 15, 1359-1367 (1997).
- Cho, R.J. *et al.* A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell* 2, 65-73 (1998).
- Gray, N.S. *et al.* Exploiting chemical libraries, structure, and genomics in the search for kinase inhibitors. *Science* 281, 533-538 (1998).
- Cardenas, M.E., Lorenz, M., Hemenway, C. & Heitman, J. Yeast as model T cells. *Perspect. Drug Discovery Design* 2, 103-126 (1994).
- Klee, C.B., Ren, H. & Wang, X. Regulation of the calmodulin-stimulated protein phosphatase, calcineurin. *J. Biol. Chem.* 273, 13367-13370 (1998).
- Tanida, I., Hasegawa, A., Iida, H., Ohya, Y. & Anraku, Y. Cooperation of calcineurin and vacuolar H(+)-ATPase in intracellular Ca^{2+} homeostasis of yeast cells. *J. Biol. Chem.* 270, 10113-10119 (1995).
- Moser, M.J., Geiser, J.R. & Davis, T.N. Ca^{2+} -calmodulin promotes survival of pheromone-induced growth arrest by activation of calcineurin and Ca^{2+} -calmodulin-dependent protein kinase. *Mol. Cell. Biol.* 16, 4824-4831 (1996).
- Mizunuma, M., Hirata, D., Miyahara, K., Tsuchiya, E. & Miyakawa, T. Role of calcineurin and Mpk1 in regulating the onset of mitosis in budding yeast. *Nature* 392, 303-306 (1998).
- Yazdanbakhsh, K., Choi, J.W., Li, Y., Lau, L.F. & Choi, Y. Cyclosporin A blocks apoptosis by inhibiting the DNA binding activity of the transcription factor Nur77. *Proc. Natl. Acad. Sci. USA* 92, 437-441 (1995).
- Molkentin, J.D. *et al.* A calcineurin-dependent transcriptional pathway for cardiac hypertrophy. *Cell* 93, 215-228 (1998).
- Mansuy, I.M., Mayford, M., Jacob, B., Kandel, E.R. & Bach, M.E. Restricted and regulated overexpression reveals calcineurin as a key component in the transition from short-term to long-term memory. *Cell* 92, 39-49 (1998).
- Schreiber, S.L. & Crabtree, G.R. The mechanism of action of cyclosporin A and FK506. *Immunol. Today* 13, 136-142 (1992).
- Cyert, M.S., Kunisawa, R., Kaim, D. & Thorner, J. Yeast has homologs (CNA1 and CNA2 gene products) of mammalian calcineurin, a calmodulin-regulated phosphoprotein phosphatase. *Proc. Natl. Acad. Sci. USA* 88, 7376-7380 (1991).
- Jones, E.W. & Fink, G.R. in *The Molecular Biology of the Yeast Saccharomyces: Metabolism and Gene Expression* (eds. Strathern, J.N., Jones, E.W. & Broach, J.R.) 181-299 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 1982).
- Hinnebusch, A. Translational regulation of yeast GCN4. *J. Biol. Chem.* 272, 21661-21664 (1997).
- Hinnebusch, A.G. in *The Molecular and Cellular Biology of the Yeast*

- Saccharomyces: Gene Expression*. (eds. Jones, E.W., Pringle, J.R. & Broach, J.R.) 319-414 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 1992).
25. Heitman, J. *et al.* The immunosuppressant FK506 inhibits amino acid import in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* 13, 5010-5019 (1993).
 26. Balzi, E. & Goffeau, A. Yeast multidrug resistance: the PDR network. *J. Bioenerg. Biomembr.* 27, 71-76 (1995).
 27. Egner, R., Rosenthal, F.E., Kralli, A., Sanglard, D. & Kuchler, K. Genetic separation of FK506 susceptibility and drug transport in the yeast Pdr5 ATP-binding cassette multidrug resistance transporter *Mol. Biol. Cell* 9, 523-543 (1998).
 28. Barkai, N. & Leibler, S. Robustness in simple biochemical networks. *Nature* 387, 913-917 (1997).
 29. Schiestl, R.H., Manivasakam, P., Woods, R.A. & Gietz, R.D. Introducing DNA into yeast by transformation. *Methods: A companion to Methods in Enzymology* 5, 79-85 (1993).
 30. Wach, A., Brachat, A., Pohlmann, R. & Philippsen, P. New heterologous modules for classical or PCR-based gene disruptions in *Saccharomyces cerevisiae*. *Yeast* 10, 1793-1808 (1994).
 31. Garrett-Engle, P., Moilanen, B. & Cyert, M.S. Calcineurin, the Ca²⁺/calmodulin-dependent protein phosphatase, is essential in yeast mutants with cell integrity defects and in mutants that lack a functional vacuolar H⁺-ATPase. *Mol. Cell. Biol.* 15, 4103-4114 (1995).
 32. Ausubel, F.M. *et al.* in *Current Protocols in Molecular Biology* 13.12.1-13.12.5 (eds. Ausubel, F.M., *et al.*) (John Wiley & Sons, New York, 1993).
 33. Bulmer, M.G. in *Principles of Statistics* 224-225 (Dover Publications, New York, 1979).
 34. Sikorski, R.S. & Hieter, P. A system of shuttle vectors and yeast host strains designated for efficient manipulation of DNA in *Saccharomyces cerevisiae*. *Genetics* 122, 19-27 (1989).
 35. Brachmann, C.B. *et al.* Designer deletion strains derived from *Saccharomyces cerevisiae* S288C: A useful set of strains and plasmids for PCR-mediated gene disruption and other applications. *Yeast* 14, 115-132 (1998).

REPORTS

co mosaic viral RNA was obtained by phenol and chloroform extractions of the virus and precipitated from ethanol. CA-NC assembly reactions in the presence of noncognate RNAs were identical to those given in (9). In the absence of RNA, CA-NC cones formed under the following conditions: 300 μ M CA-NC, 1 M NaCl, and 50 mM Tris-HCl (pH 8.0) at 37°C for 60 min. In the absence of exogenous RNA, neither cones nor cylinders formed at concentrations of 0.5 M NaCl or below. Absorption spectra demonstrated that our CA-NC preparations were not contaminated with *Escherichia coli* RNA (estimated lower detection limit was \sim 1 base/protein molecule). To control for even lower levels of RNA contamination, we preincubated the CA-NC protein with 0.5 mg/ml ribonuclease A (Type 1-AS, 54 Kunitz U/mg, Sigma) for 1 hour at 4°C, which then formed cones normally.

13. V. Y. Klishko, data not shown.

14. M. Ge and K. Sattler, *Chem. Phys. Lett.* 220, 192 (1994).

15. A. Krishnan et al., *Nature* 388, 451 (1997).

16. L. B. Kong et al., *J. Virol.* 72, 4403 (1998).

17. Assembly mixtures were deposited on holey carbon grids, blotted briefly with filter paper, plunged into liquid ethane, and transferred to liquid nitrogen. Frozen grids were transferred to a Philips 420 TEM equipped with a Gatan cold stage system, and images of particles in vitreous ice were recorded under low dose conditions at 36,000 \times magnification and \sim 1.6- μ m defocus.

18. J. T. Finch, data not shown.

19. R. A. Crowther, *Proceedings of the Third John Innes Symposium* (1976), pp. 15-25; E. Kellenberger, M. Häner, M. Wurtz, *Ultramicroscopy* 9, 139 (1982); J. Seymour and D. J. DeRosier, *J. Microsc.* 148, 195 (1987).

20. M. V. Nermut, C. Grief, S. Hashmi, D. J. Hockley, *AIDS Res. Hum. Retroviruses* 9, 929 (1993); M. V. Nermut et al., *Virology* 198, 288 (1994); E. Barklis, J. McDermott, S. Wilkens, S. Fuller, D. Thompson, *J. Biol. Chem.* 273, 7177 (1998); E. Barklis et al., *EMBO J.* 16, 1199 (1997); M. Yeager, E. M. Wilson-Kubalek, S. G. Weiner, P. O. Brown, A. Rein, *Proc. Natl. Acad. Sci. U.S.A.* 95, 7299 (1998).

21. J. T. Finch et al., unpublished observations.

22. V. M. Vogt, in (2), pp. 27-70.

23. M. A. McClure, M. S. Johnson, D.-F. Feng, R. F. Doolittle, *Proc. Natl. Acad. Sci. U.S.A.* 85, 2469-2473 (1988).

24. Single-letter abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.

25. We thank C. Hill for very helpful discussions on the relationship between viral cores and fullerene cones, D. Hobbs for refining the ChemDraw3D images of cones, G. Stubbs for a gift of tobacco mosaic virus, J. McCutcheon for the plasmid used to prepare ribosomal RNA, and K. Albertine and N. Chandler of the University of Utah Shared Electron Microscopy facility for their support and encouragement. Supported by grants from NIH and from the Huntsman Cancer Institute (to W.L.S.).

29 September 1998; accepted 17 November 1998

The Transcriptional Program in the Response of Human Fibroblasts to Serum

Vishwanath R. Iyer, Michael B. Eisen, Douglas T. Ross, Greg Schuler, Troy Moore, Jeffrey C. F. Lee, Jeffrey M. Trent, Louis M. Staudt, James Hudson Jr., Mark S. Boguski, Deval Lashkari, Dari Shalon, David Botstein, Patrick O. Brown*

The temporal program of gene expression during a model physiological response of human cells, the response of fibroblasts to serum, was explored with a complementary DNA microarray representing about 8600 different human genes. Genes could be clustered into groups on the basis of their temporal patterns of expression in this program. Many features of the transcriptional program appeared to be related to the physiology of wound repair, suggesting that fibroblasts play a larger and richer role in this complex multicellular response than had previously been appreciated.

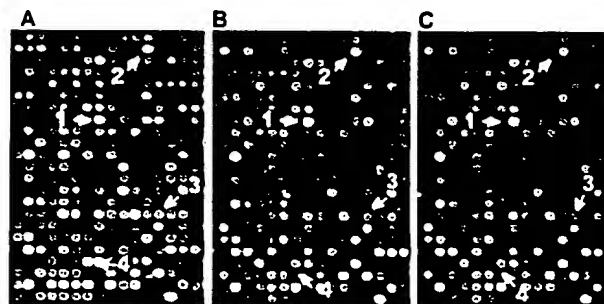
The response of mammalian fibroblasts to serum has been used as a model for studying growth control and cell cycle progression (1). Normal human fibroblasts require growth factors for proliferation in culture; these growth factors are usually provided by fetal

bovine serum (FBS). In the absence of growth factors, fibroblasts enter a nondividing state, termed G_0 , characterized by low

metabolic activity. Addition of FBS or purified growth factors induces proliferation of the fibroblasts; the changes in gene expression that accompany this proliferative response have been the subject of many studies, and the responses of dozens of genes to serum have been characterized.

We took a fresh look at the response of human fibroblasts to serum, using cDNA microarrays representing about 8600 distinct human genes to observe the temporal program of transcription that underlies this response. Primary cultured fibroblasts from human neonatal foreskin were induced to enter a quiescent state by serum deprivation for 48 hours and then stimulated by addition of medium containing 10% FBS (2). DNA microarray hybridization was used to measure the temporal changes in mRNA levels of 8613 human genes (3) at 12 times, ranging from 15 min to 24 hours after serum stimulation. The cDNA made from purified mRNA from each sample was labeled with the fluorescent dye Cy5 and mixed with a common reference probe consisting of cDNA made from purified mRNA from the quiescent

Fig. 1. The same section of the microarray is shown for three independent hybridizations comparing RNA isolated at the 8-hour time point after serum treatment to RNA from serum-deprived cells. Each microarray contained 9996 elements, including 9804 human cDNAs, representing 8613 different genes. mRNA from serum-deprived cells was used to prepare cDNA labeled with



Cy3-deoxyuridine triphosphate (dUTP), and mRNA harvested from cells at different times after serum stimulation was used to prepare cDNA labeled with Cy5-dUTP. The two cDNA probes were mixed and simultaneously hybridized to the microarray. The image of the subsequent scan shows genes whose mRNAs are more abundant in the serum-deprived fibroblasts (that is, suppressed by serum treatment) as green spots and genes whose mRNAs are more abundant in the serum-treated fibroblasts as red spots. Yellow spots represent genes whose expression does not vary substantially between the two samples. The arrows indicate the spots representing the following genes: 1, protein disulfide isomerase-related protein P5; 2, IL-8 precursor; 3, EST AA057170; and 4, vascular endothelial growth factor.

V. R. Iyer and D. T. Ross, Department of Biochemistry, Stanford University School of Medicine, Stanford CA 94305, USA. M. B. Eisen and D. Botstein, Department of Genetics, Stanford University School of Medicine, Stanford CA 94305, USA. G. Schuler and M. S. Boguski, National Center for Biotechnology Information, Bethesda MD 20894, USA. T. Moore and J. Hudson Jr., Research Genetics, Huntsville, AL 35801, USA. J. C. F. Lee, D. Lashkari, D. Shalon, Incyte Pharmaceuticals, Fremont, CA 94555, USA. J. M. Trent, Laboratory of Cancer Genetics, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA. L. M. Staudt, Metabolism Branch, Division of Clinical Sciences, National Cancer Institute, Bethesda, MD 20892, USA. P. O. Brown, Department of Biochemistry and Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford CA 94305, USA.

*To whom correspondence should be addressed. E-mail: pbrown@cmgm.stanford.edu

REPORTS

culture (time zero) labeled with a second fluorescent dye, Cy3 (4). The color images of the hybridization results (Fig. 1) were made by representing the Cy3 fluorescent image as green and the Cy5 fluorescent image as red and merging the two color images.

Diverse temporal profiles of gene expression could be seen among the 8613 genes sur-

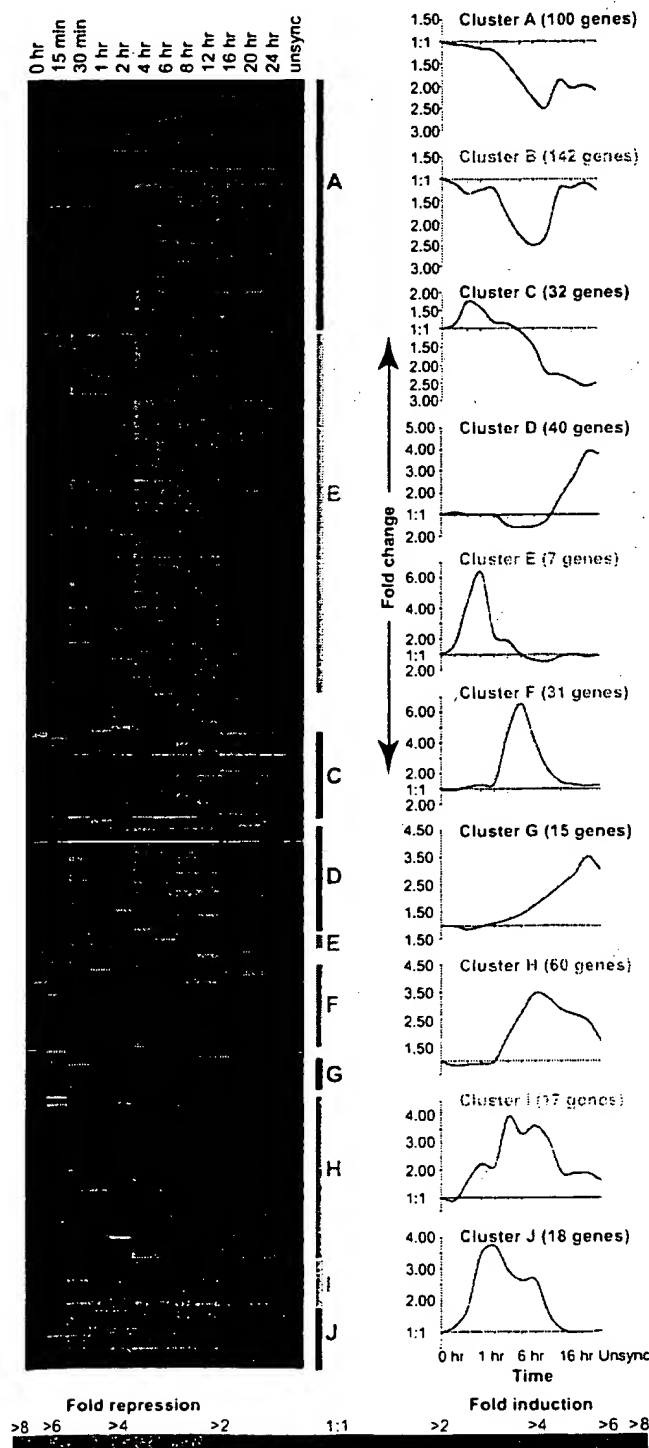
veyed in this experiment (Fig. 2); many of these genes (about half) were unnamed expressed sequence tags (ESTs) (5). Although diverse patterns of expression were observed, the orderly choreography of the expression program became apparent when the results were analyzed by a clustering and display method developed in our laboratory for analyzing genome-wide

gene expression data (6). An example of such an analysis, here applied to a subset of 517 genes whose expression changed substantially in response to serum (7), is shown in Fig. 2. The entire detailed data set underlying Fig. 2 is available as a tab-delimited table (in cluster order) at the Science Web site (www.sciencemag.org/feature/data/984559.shl). In addition, the entire, larger data set for the complete set of genes analyzed in this experiment can be found at a Web site maintained by our laboratory (genome-www.stanford.edu/serum) (8).

One measure of the reliability of the changes we observed is inherent in the expression profiles of the genes. For most genes whose expression levels changed, we could see a gradual change over a few time points, which thus effectively provided independent measurements for almost all of the observations. An additional check was provided by the inclusion of duplicate and, in a few cases, multiple array elements representing the same gene for about 5% of the genes included in this microarray. In addition, three independent hybridizations to different microarrays with mRNA samples from cells harvested 8 hours after serum addition showed good correlation (Fig. 1). As an independent test, we measured the expression levels of several genes using the TaqMan 5' nuclease fluorogenic quantitative polymerase chain reaction (PCR) assay (9). The expression profiles of the genes, as measured by these two independent methods, were very similar (Fig. 3) (10).

The transcriptional response of fibroblasts to serum was extremely rapid. The immediate response to serum stimulation was dominated by genes that encode transcription factors and other proteins involved in signal transduction. The mRNAs for several genes [including c-FOS, JUN B, and mitogen-activated protein (MAP) kinase phosphatase-1 (MKP1)] were detectably induced within 15 min after serum stimulation (Fig. 4, A and B). Fifteen of the genes that were observed to be induced by serum encode known or suspected regulators of transcription (Fig. 4B). All but one were immediately induced—their induction was not inhibited by cycloheximide (11). This class of genes could be distinguished into those whose induction was transient (Fig. 2, cluster E) and those whose mRNA levels remained induced for much longer (Fig. 2, clusters I and J). Some features of the immediate response appeared to be directed at adaptation to the initiating signals. We observed a marked induction of mRNA encoding MKP1, a dual-specificity phosphatase that modulates the activity of the ERK1 and ERK2 MAP kinases (12). The coincidence of the peak of expression of genes in cluster E (Fig. 2) with that of MKP1 (Fig. 4A) suggests the possibility

Fig. 2. Cluster image showing the different classes of gene expression profiles. Five hundred seventeen genes whose mRNA levels changed in response to serum stimulation were selected (7). This subset of genes was clustered hierarchically into groups on the basis of the similarity of their expression profiles by the procedure of Eisen *et al.* (6). The expression pattern of each gene in this set is displayed here as a horizontal strip. For each gene, the ratio of mRNA levels in fibroblasts at the indicated time after serum stimulation ("unsync" denotes exponentially growing cells) to its level in the serum-deprived (time zero) fibroblasts is represented by a color, according to the color scale at the bottom. The graphs show the average expression profiles for the genes in the corresponding "cluster" (indicated by the letters A to J and color coding). In every case examined, when a gene was represented by more than one array element, the multiple representations in this set were seen to have identical or very similar expression profiles, and the profiles corresponding to these independent measurements clustered either adjacent or very close to each other, pointing to the robustness of the clustering algorithm in grouping genes with very similar patterns of expression.



REPORTS

that continued activity of the MAP kinase pathway is required to maintain induction of these genes but not of those with sustained expression (clusters I and J). The gene encoding a second member of the dual-specificity MAP kinase phosphatase family, known as dual-specificity protein phosphatase 6/pyst2, was induced later, at about 4 hours after serum stimulation. Genes encoding diverse other proteins with roles in signal transduction, ranging from cell-surface receptors [for example, the sphingosine 1-phosphate receptor (EDG-1), the vascular endothelial growth factor receptor, and the type II BMP receptor] to regulators of G-protein signaling (for example, NET1/p115 rho GEF) to DNA-binding transcription factors, were induced by serum (Fig. 4A).

The reprogramming of the regulatory circuits in response to serum involved not only induction of transcription factors but also reduced expression of many transcriptional regulators—some of which may play roles in maintaining the cells in G_0 or in priming them to react to wounding (Fig. 4C). Perhaps as a consequence of the historical focus on genes induced by serum stimulation of fibroblasts, the set of transcription factors whose expression diminished upon serum stimulation has been less well characterized.

Genes known or likely to be involved in controlling and mediating the proliferative response showed distinctive patterns of regulation. Several genes whose products inhibit progression of the cell-division cycle, such as p27 Kip1, p57 Kip2, and p18, were expressed in the quiescent fibroblasts and down-regulated before the onset of cell division. The nadir in the mRNA levels for these genes occurred between 6 and 12 hours after serum stimulation (Fig. 5A), coincident with the passage of the fibroblasts through G_1 . The levels of the transcript encoding the WEE1-like protein kinase, which is believed to inhibit mitosis by phosphorylation of Cdc2, diminished between 4 and 8 to 12 hours after serum addition (Fig. 5A), well

before the onset of M phase at around 16 hours, raising the possibility of an additional role for Wee1 in an earlier stage of the cell cycle or in regulating the G_0 to G_1 transition. Several genes induced in the first few hours after serum stimulation, such as the helix-loop-helix proteins ID2 and ID3 and EST AA016305, a gene with homology to G_1 -S cyclins, are candidates for roles in promoting the exit from G_0 .

Genes involved in mediating progression through the cell cycle were characterized by a distinctive pattern of expression (Fig. 2, cluster D), reflecting the coincidence of their expression with the reentry of the stimulated fibroblasts into the cell-division cycle. The stimulated fibroblasts replicated their DNA about 16 hours after serum treatment. This timing was reflected by the induction of mRNA encoding both subunits of ribonucleotide reductase and PCNA, the processivity factor for DNA polymerase epsilon and delta. Cyclin A, Cyclin B1, Cdc2, and CDC28 kinase, regulators of passage through the S phase and the transition from G_2 to M phase, were induced at about 16 to 20 hours after serum addition. The kinase in the Cyclin B1-CDK pair needs to be activated by phosphorylation. The gene encoding Cyclin-dependent kinase 7 (CDK7; a homolog of *Xenopus* MO15 cdk-activating kinase) was induced in parallel with the Cdc2 and Cdc28 kinases (Fig. 5A), suggesting a potential role for CDK7 in mediating M phase. DNA topoisomerase II α , required for chromosome segregation at mitosis; Mad2, a component of the spindle checkpoint that prevents completion of mitosis (anaphase) if chromosomes are not attached to the spindle; and the kinetochore protein CENP-F all showed a similar expression profile.

In the hours after the serum stimulus, one of the most striking features of the unfolding transcriptional program was the appearance of numerous genes with known roles in processes relevant to the physiology of wound healing.

These included both genes involved in the direct role played by fibroblasts in remodeling of the clot and the extracellular matrix and, more notably, genes encoding proteins involved in intercellular signaling (Fig. 5). Genes induced in this program encode products that can (i) participate in the dynamic process of clotting, clot dissolution, and remodeling and perhaps contribute to hemostasis by promoting local vasoconstriction (for example, endothelin-1); (ii) promote chemotaxis and activation of neutrophils (for example, COX2) and recruitment and extravasation of monocytes and macrophages (for example, MCP1); (iii) promote chemotaxis and activation of T lymphocytes [for example, interleukin-8 (IL-8)] and B lymphocytes (for example, ICAM-1), thus providing both innate and antigen-specific defenses against wound infection and recruiting the phagocytic cells that will be required to clear out the debris during remodeling of the wound; (iv) promote angiogenesis and neovascularization (for example, VEGF) through newly forming tissue; (v) promote migration and proliferation of fibroblasts (for example, CTGF) and their differentiation into myofibroblasts (for example, Vimentin); and (vi) promote migration and proliferation of keratinocytes, leading to reepithelialization of the wound (for example, FGF7), and promote proliferation of melanocytes, perhaps contributing to wound hyperpigmentation (for example, FGF2).

Coordinated regulation of groups of genes whose products act at different steps in a common process was a recurring theme. For example, Furin, a prohormone-processing protease required for one of the processing steps in the generation of active endothelin, was induced in parallel with induction of the gene encoding the precursor of endothelin-1 (Fig. 5E) (13). Conversely, expression of CALLA/CD10, a membrane metalloprotease that degrades endothelin-1 and other peptide mediators of acute inflammation, was re-

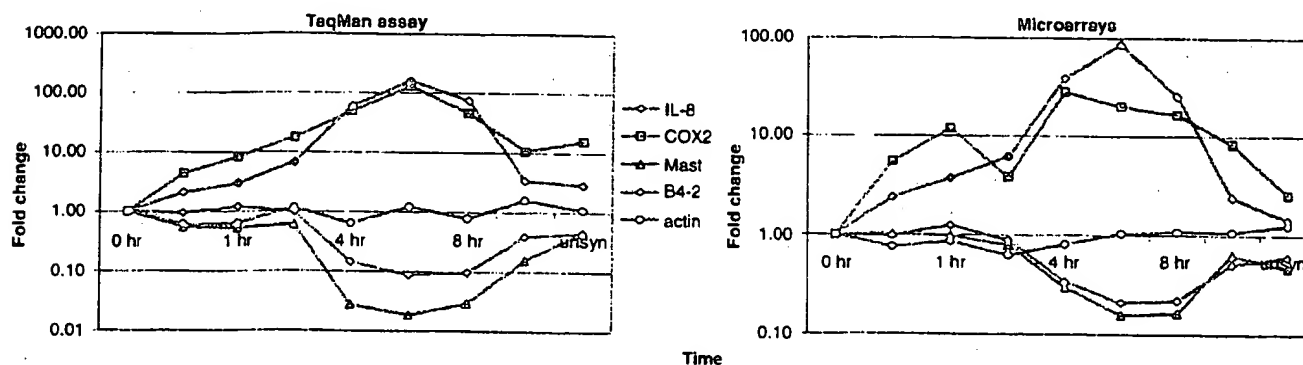


Fig. 3. Independent verification of microarray quantitation. Relative mRNA levels of the indicated genes (Mast, mast/stem cell growth factor receptor) were measured with the TaqMan 5' nuclease fluorogenic quantitative PCR assay (9) (left) in the same samples that were used to prepare probes for microarray hybridizations (right). Data from the TaqMan analysis were

normalized to mRNA concentrations and plotted relative to the level at time zero, so that the results could be compared with those from the microarray hybridizations. In general, quantitation with the two methods gave very similar results (10).

REPORTS

duced. A second example is provided by a set of five genes involved in the biosynthesis of cholesterol (Fig. 5I). The mRNAs encoding each of these enzymes showed sharply diminished expression beginning 4 to 6 hours after serum stimulation of fibroblasts. A likely explanation for the coordinated down-regulation of the cholesterol biosynthetic pathway is that serum provides cholesterol to fibroblasts through low-density lipoproteins, whereas in the absence of the cholesterol provided by serum, endogenous cholesterol biosynthesis in fibroblasts is required.

Many of the previously studied genes that we observed to be regulated in this program have no recognized role in any aspect of wound healing or fibroblast proliferation. Their identification in this study may therefore point to previously unknown aspects of these processes. A few selected genes in this group are shown in Fig. 5H. The stanniocalcin gene, for example (Fig. 5H), encodes a secreted protein without a clearly identified function in human cells (14, 15). Its induction in serum-stimulated fibro-

blasts suggests the possibility that it may play a role in the wound-healing process, perhaps serving as a signal in mediating inflammation or angiogenesis.

One of the most important results of this exploration was the discovery of over 200 previously unknown genes whose expression was regulated in specific temporal patterns during the response of fibroblasts to serum. For example, 13 of the 40 genes in cluster D (Fig. 2) have descriptive names that reflect their putative function. Nine of these 13 genes (69%) encode proteins that play roles in cell cycle progression, particularly in DNA replication and the G₂-M transition. This enrichment for cell cycle-related genes suggests that some of the

unnamed genes in this cluster—for example, EST W79311 and EST R13146, neither of which have sequence similarity to previously characterized genes—may represent previously unknown genes involved in this part of the cell cycle. Similarly, a remarkable fraction of genes that were grouped into cluster F on the basis of their expression profiles encoded proteins involved in intercellular signaling (Fig. 2), suggesting that a similar role should be considered for the many unnamed genes in this cluster. A disproportionately large fraction of the genes whose transcription diminished upon serum stimulation were unnamed ESTs.

Our intention was to use this experiment as a model to study the control of the transition

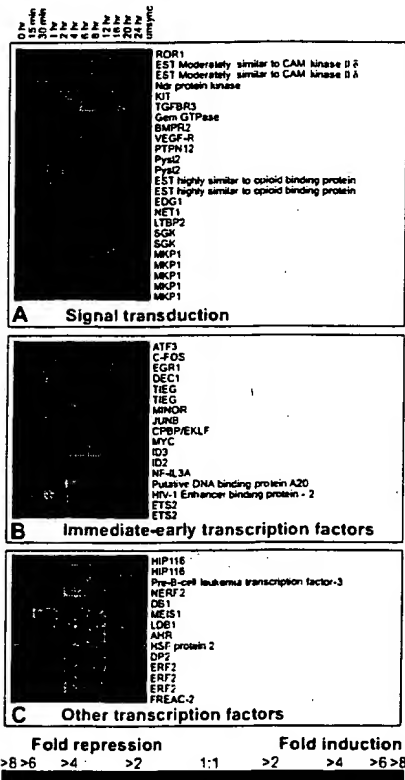


Fig. 4. "Reprogramming" of fibroblasts. Expression profiles of genes whose function is likely to play a role in the reprogramming phase of the response are shown with the same representation as in Fig. 2. In the cases in which a gene was represented by more than one element in the microarray, all measurements are shown. The genes were grouped into categories on the basis of our knowledge of their most likely role. Some genes with pleiotropic roles were included in more than one category.

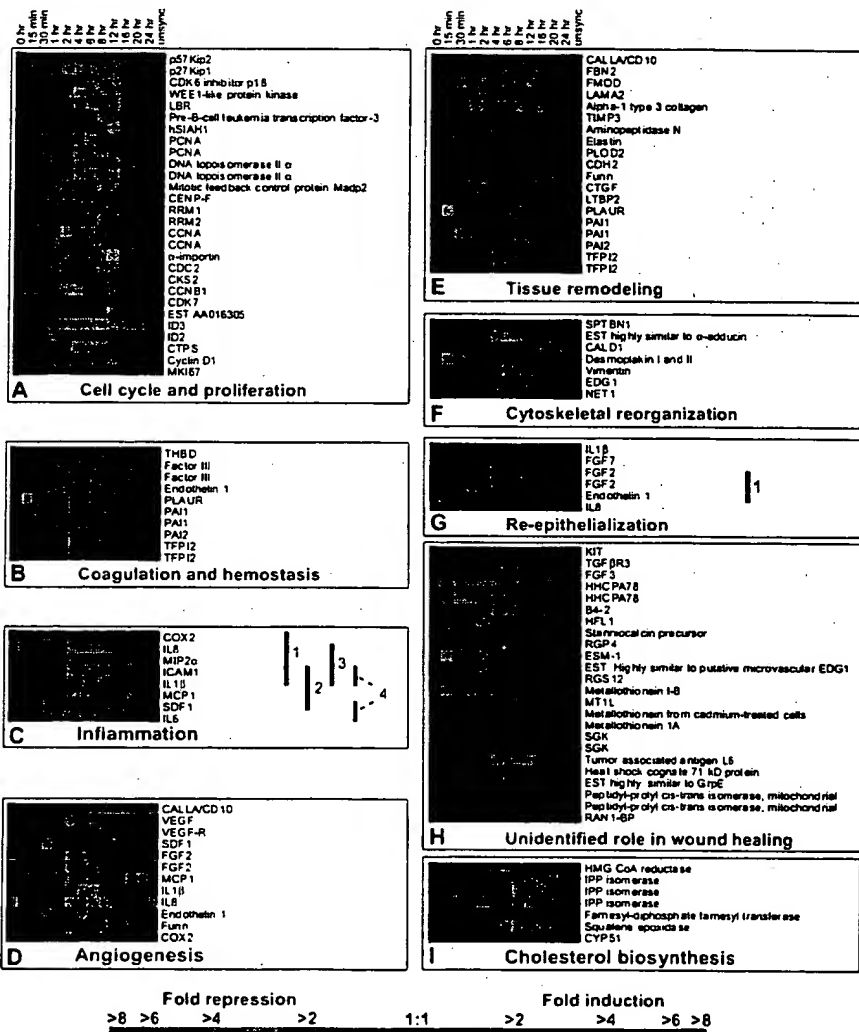


Fig. 5. The transcriptional response to serum suggests a multifaceted role for fibroblasts in the physiology of wound healing. The features of the transcriptional program of fibroblasts in response to serum stimulation that appear to be related to various aspects of the wound-healing process and fibroblast proliferation are shown with the same convention for representing changes in transcript levels as was used in Figs. 2 and 4. (A) Cell cycle and proliferation, (B) coagulation and hemostasis, (C) inflammation, (D) angiogenesis, (E) tissue remodeling, (F) cytoskeletal reorganization, (G) reepithelialization, (H) unidentified role in wound healing, and (I) cholesterol biosynthesis. The numbers in (C) and (G) refer to genes whose products serve as signals to neutrophils (C1), monocytes and macrophages (C2), T lymphocytes (C3), B lymphocytes (C4), and melanocytes (G1).

REPORTS

from G₀ to a proliferating state. However, one of the defining characteristics of genome-scale expression profiling experiments is that the examination of so many diverse genes opens a window on all the processes that actually occur and not merely the single process one intended to observe. Serum, the soluble fraction of clotted blood, is normally encountered by cells in vivo in the context of a wound. Indeed, the expression program that we observed in response to serum suggests that fibroblasts are programmed to interpret the abrupt exposure to serum not as a general mitogenic stimulus but as a specific physiological signal, signifying a wound. The proliferative response that we originally intended to study appeared to be part of a larger physiological response of fibroblasts to a wound. Other features of the transcriptional response to serum suggest that the fibroblast is an active participant in a conversation among the diverse cells that work together in wound repair, interpreting, amplifying, modifying, and broadcasting signals controlling inflammation, angiogenesis, and epithelial regrowth during the response to an injury.

We recognize that these in vitro results almost certainly represent a distorted and incomplete rendering of the normal physiological response of a fibroblast to a wound. Moreover, only the responses elicited directly by exposure of fibroblasts to serum were examined. The subsequent signals from other cellular participants in the normal wound-healing process would certainly provoke further evolution of the transcriptional program in fibroblasts at the site of a wound, which this experiment cannot reveal. Nevertheless, we believe that the picture that emerged strongly suggests a much larger and richer role for the fibroblast in the orchestration of this important physiological process than had previously been suspected.

References and Notes

1. J. A. Winkles, *Prog. Nucleic Acid Res. Mol. Biol.* **58**, 41 (1998).
2. A normal human diploid fibroblast cell line derived from foreskin (ATCC CRL 2091) in passage 8 was used in these experiments. The protocol followed for growth arrest and stimulation was essentially that of (16) and (17). Cells were grown to about 60% confluence in 15-cm petri dishes in Dulbecco's minimum essential medium containing glucose (1 g/liter), the antibiotics penicillin and streptomycin, and 10% (by vol) FBS (HyClone) that had been previously heat inactivated at 56°C for 30 min. The cells were then washed three times with the same medium lacking FBS, and lower-serum medium (0.1% FBS) was added to the plates. After a 48-hour incubation, the medium was replaced with fresh medium containing 10% FBS. mRNA was isolated from several plates of cells harvested before serum stimulation; this mRNA served as the serum-starved or time-zero reference sample. Cells were harvested from batches of plates at 11 subsequent intervals (15 min, 30 min, 1, 2, 4, 6, 8, 12, 16, 20, and 24 hours) after the addition of serum. mRNA was also isolated from exponentially growing fibroblasts (not subjected to serum starvation). mRNA was isolated with the FastTrack mRNA isolation kit (Invitrogen), which involves lysis of the cells on the plate. The growth medium was removed, and the cells were quickly washed with phosphate-buffered saline at room temperature. The lysis buffer was added to the plate, transferred to tubes, and frozen in liquid nitrogen. Subsequent steps were performed according to the kit manufacturer's protocols.
3. The National Center for Biotechnology Information maintains the UniGene database as a resource for partitioning human sequences contained in GenBank into clusters representing distinct transcripts or genes (18, 19). At the time this work began, this database contained about 40,000 such clusters. We selected a subset of 10,000 of these UniGene clusters for inclusion on gene expression microarrays. UniGene clusters were included only if they contained at least one clone from the IMAGE human cDNA collection (20), so that a physical clone could easily be obtained (all IMAGE clones are available commercially from a number of vendors). We attempted to include as complete as possible a set of the "named" human genes (about 4000) and genes that appeared to be closely related to named genes in other organisms (about an additional 2000). The remaining 4000 clones were chosen from among the "anonymous" UniGene clusters on the basis of inclusion on the human transcript map (www.ncbi.nlm.nih.gov/SCIENCE96/) and the lack of apparent homology to any other genes in the selected set. A physical clone representing each of the selected genes was obtained from Research Genetics. This "10K set" is included in a more recent "15K set" described at www.nhgr.nih.gov/DIR/LCC/15K/HTML/p15Ktop.html. Of these clones, 472 are absent from the current edition of UniGene and were presumed to be distinct genes. The remainder represent 8141 distinct clusters, or human genes, in UniGene. These clones, thus presumed to represent 8613 different genes, were used to print microarrays according to methods described previously (21, 22).
4. One microgram of mRNA was used for making fluorescently labeled cDNA probes for hybridizing to the microarrays, with the protocol described previously (23). mRNA from the large batch of serum-starved cells was used to make cDNA labeled with Cy3. The Cy3-labeled cDNA from this batch of serum-starved cells served as the common reference probe in all hybridizations. mRNA samples from cells harvested immediately before serum stimulation, at intervals after serum stimulation, and from exponentially growing cells were used to make cDNA labeled with Cy5. Ten micrograms of yeast tRNA, 10 µg of polydeoxyadenylic acid, and 20 µg of human CoT1 DNA (Gibco-BRL) were added to the mixture of labeled probes in a solution containing 3× standard saline citrate (SSC) and 0.3% SDS and allowed to prehybridize at room temperature for 30 min before the probe was added to the surface of the microarray. Hybridizations, washes, and fluorescent scans were performed as described previously (23, 24). All measurements, totaling more than 180,000 differential expression measurements, were stored in a computer database for analysis and interpretation.
5. The nominal identities of a number of cDNAs (currently about 3750) on the microarray were verified by sequencing. The clones that were sequenced included many of the genes whose expression changed substantially upon serum stimulation, as well as a large number of genes whose expression did not change substantially in the course of this experiment. About 85% of the clones on the current version of this microarray that were checked by resequencing were correctly identified. In all the figures, gene names or EST numbers are given only for those genes on the microarrays whose identities were reconfirmed by resequencing. In the cases where a human gene has more than one name in the literature, we have tried to use the name that is most evocative of its presumed role in this context. The remainder of the clones have been assigned a temporary identification number (format: SID#####) and a putative identity pending sequence verification. The correct identities of these genes will be posted at our Web site (genome-www.stanford.edu/serum) as they are confirmed by resequencing.
6. M. B. Eisen, P. T. Spellman, P. O. Brown, D. Botstein, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 14863 (1998).
7. Genes were selected for this analysis if either (i) their expression level deviated from that in quiescent fibroblasts by at least a factor of 2.20 in at least two of the samples from serum-stimulated cells or (ii) the standard deviation for the set of 13 values of log₂(expression ratio) measured for the gene in this time course exceeded 0.7. In addition, observations in which the pixel-by-pixel correlation coefficients for the Cy3 and Cy5 fluorescence signals measured in a given array element were less than 0.6 were excluded. This selection criteria yielded a computationally manageable number of genes while minimizing the number of genes that were included because of noise in the data.
8. A more complete analysis and interpretation of the results of this experiment, as well as a searchable database, can be found at genome-www.stanford.edu/serum.
9. K. J. Livak, S. J. Flood, J. Marmaro, W. Giusti, K. Deetz, *PCR Methods. Appl.* **4**, 357 (1995).
10. The apparent dip in the profile of COX2 at the 2-hour time point in the microarray data appears to result from a localized area of low intensity on the corresponding array scan resulting in an underestimation of the expression ratio. The expression ratios measured for mast/stem cell growth factor receptor are somewhat lower in the microarray data. This discrepancy is probably a consequence of the conservative background subtraction method used for quantitating the signal intensities on the array scans (23). The sequences of the PCR primer pairs (5' to 3') that were used are as follows: COX2, CCGTGGCTCTCTT-CGCAG and CTAAGTTCTTAGCACTCCTTGGCA; IL-8, CGATGCTGGAGCTGTATC and CCATGGTTTC-ACCAAGATC; mast/stem cell factor receptor, ACA-GAAGCCCTGCTACACC and GAGCTCGGAGGAG-GAAG; B4-2, AAACCCCTTCAGGAAGAG and CC-ATGAACAAGCTGGCCAT; and actin, ACTACTCCGTGTCGATCCGC and GCTGATCCACATCTGCTGA.
11. V. R. Iyer et al., unpublished data. The gene expression data for the early time points in the presence of cycloheximide will be available at our Web site (genome-www.stanford.edu/serum).
12. T. Hunter, *Cell* **80**, 225 (1995).
13. J. Leppaluoto and H. Ruskoaho, *Ann. Med.* **24**, 153 (1992).
14. A. C. Chang et al., *Mol. Cell. Endocrinol.* **112**, 241 (1995).
15. K. L. Madsen et al., *Am. J. Physiol.* **274**, G96 (1998).
16. W. Krek and J. A. DeCaprio, *Methods Enzymol.* **254**, 114 (1995).
17. R. A. Tobey, J. G. Valdez, H. A. Crissman, *Exp. Cell Res.* **179**, 400 (1988).
18. M. S. Boguski and G. D. Schuler, *Nature Genet.* **10**, 369 (1995).
19. G. D. Schuler, *J. Mol. Med.* **75**, 694 (1997).
20. C. Lennon, C. Auffray, M. Polymeropoulos, M. B. Soares, *Genomics* **33**, 151 (1996).
21. IMAGE clones were amplified by PCR in 96-well format with amino-linked primers at the 5' end. Purified PCR products were suspended at a concentration of ~0.5 mg/ml in 3× SSC, and ~5 ng of each product was arrayed onto coated glass by means of procedures similar to those described previously (22). A total of 9996 elements were arrayed onto an area of 1.8 cm by 1.8 cm with the elements spaced 175 µm apart. The microarrays were then postprocessed to fix the DNA to the glass surface before hybridization with a procedure similar to previously described methods (22).
22. M. Schena, D. Shalon, R. W. Davis, P. O. Brown, *Science* **270**, 467 (1995).
23. J. L. DeRisi, V. R. Iyer, P. O. Brown, *ibid.* **278**, 680 (1997).
24. J. DeRisi et al., *Nature Genet.* **14**, 457 (1996).
25. We thank E. Chung for help with sequencing, A. Alizadeh for help with sequence verification, K. Ranade for advice on the TaqMan assay, and J. DeRisi and other members of the P.O.B. and D.B. labs for discussions. Supported by a grant from the National Human Genome Research Institute (NHGRI) (HG00450) and the National Cancer Institute (NIH CA 77097). V.R.I. was supported in part by an Institutional Training Grant in Genome Sciences (T32 HG00044) from the NHGRI. M.B.E. is an Alfred E. Sloan Foundation Postdoctoral Fellow in Computational Molecular Biology, and D.T.R. is a Walter and Idun Berry Fellow. P.O.B. is an Associate Investigator of the Howard Hughes Medical Institute.

13 August 1998; accepted 13 November 1998

Systematic variation in gene expression patterns in human cancer cell lines

Douglas T. Ross¹, Uwe Scherf², Michael B. Eisen², Charles M. Perou², Christian Rees², Paul Spellman², Vishwanath Iyer¹, Stefanie S. Jeffrey³, Matt Van de Rijn⁴, Mark Waltham⁵, Alexander Pergamenschikov², Jeffrey C.F. Lee⁶, Deval Lashkari⁷, Dari Shalon⁶, Timothy G. Myers⁸, John N. Weinstein⁵, David Botstein² & Patrick O. Brown^{1,9}

We used cDNA microarrays to explore the variation in expression of approximately 8,000 unique genes among the 60 cell lines used in the National Cancer Institute's screen for anti-cancer drugs. Classification of the cell lines based solely on the observed patterns of gene expression revealed a correspondence to the ostensible origins of the tumours from which the cell lines were derived. The consistent relationship between the gene expression patterns and the tissue of origin allowed us to recognize outliers whose previous classification appeared incorrect. Specific features of the gene expression patterns appeared to be related to physiological properties of the cell lines, such as their doubling time in culture, drug metabolism or the interferon response. Comparison of gene expression patterns in the cell lines to those observed in normal breast tissue or in breast tumour specimens revealed features of the expression patterns in the tumours that had recognizable counterparts in specific cell lines, reflecting the tumour, stromal and inflammatory components of the tumour tissue. These results provided a novel molecular characterization of this important group of human cell lines and their relationships to tumours *in vivo*.

Introduction

Cell lines derived from human tumours have been extensively used as experimental models of neoplastic disease. Although such cell lines differ from both normal and cancerous tissue, the inaccessibility of human tumours and normal tissue makes it likely that such cell lines will continue to be used as experimental models for the foreseeable future. The National Cancer Institute's Developmental Therapeutics Program (DTP) has carried out intensive studies of 60 cancer cell lines (the NCI60) derived from tumours from a variety of tissues and organs¹⁻⁴. The DTP has assessed many molecular features of the cells related to cancer and chemotherapeutic sensitivity, and has measured the sensitivities of these 60 cell lines to more than 70,000 different chemical compounds, including all common chemotherapeutics (<http://dtp.nci.nih.gov>). A previous analysis of these data revealed a connection between the pattern of activity of a drug and its method of action. In particular, there was a tendency for groups of drugs with similar patterns of activity to have related methods of action^{3,5-7}.

We used DNA microarrays to survey the variation in abundance of approximately 8,000 distinct human transcripts in these 60 cell lines. Because of the logical connection between the function of a gene and its pattern of expression, the correlation of gene expression patterns with the variation in the phenotype of the cell can begin the process by which the function of a gene can be inferred. Similarly, the patterns of expression of known genes can

reveal novel phenotypic aspects of the cells and tissues studied⁸⁻¹⁰. Here we present an analysis of the observed patterns of gene expression and their relationship to phenotypic properties of the 60 cell lines. The accompanying report¹¹ explores the relationship between the gene expression patterns and the drug sensitivity profiles measured by the DTP. The assessment of gene expression patterns in a multitude of cell and tissue types, such as the diverse set of cell lines we studied here, under diverse conditions *in vitro* and *in vivo*, should lead to increasingly detailed maps of the human gene expression program and provide clues as to the physiological roles of uncharacterized genes¹¹⁻¹⁶. The databases, plus tools for analysis and visualization of the data, are available (<http://genome-www.stanford.edu/nci60> and <http://discover.nci.nih.gov>).

Results

We studied gene expression in the 60 cell lines using DNA microarrays prepared by robotically spotting 9,703 human cDNAs on glass microscope slides^{17,18}. The cDNAs included approximately 8,000 different genes: approximately 3,700 represented previously characterized human proteins, an additional 1,900 had homologues in other organisms and the remaining 2,400 were identified only by ESTs. Due to ambiguity of the identity of the cDNA clones used in these studies, we estimated that approximately 80% of the genes in these experiments were correctly identified. The identities of approximately 3,000 cDNAs

Departments of ¹Biochemistry, ²Genetics, ³Surgery and ⁴Pathology, Stanford University School of Medicine, Stanford, California, USA. ⁵Laboratory of Molecular Pharmacology, Division of Basic Sciences, National Cancer Institute, National Institutes of Health, Bethesda, Maryland, USA. ⁶Incyte Pharmaceuticals, Fremont, California, USA. ⁷Genometrix Inc., The Woodlands, Texas, USA. ⁸Information Technology Branch, Developmental Therapeutics Program, Division of Cancer Treatment and Diagnosis, National Cancer Institute, National Institutes of Health, Rockville, Maryland, USA. ⁹Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, California, USA. Correspondence should be addressed to P.O.B. (e-mail: pbrown@cmgm.stanford.edu) or J.N.W. (e-mail: Weinstein@dtipax2.ncifcrf.gov).

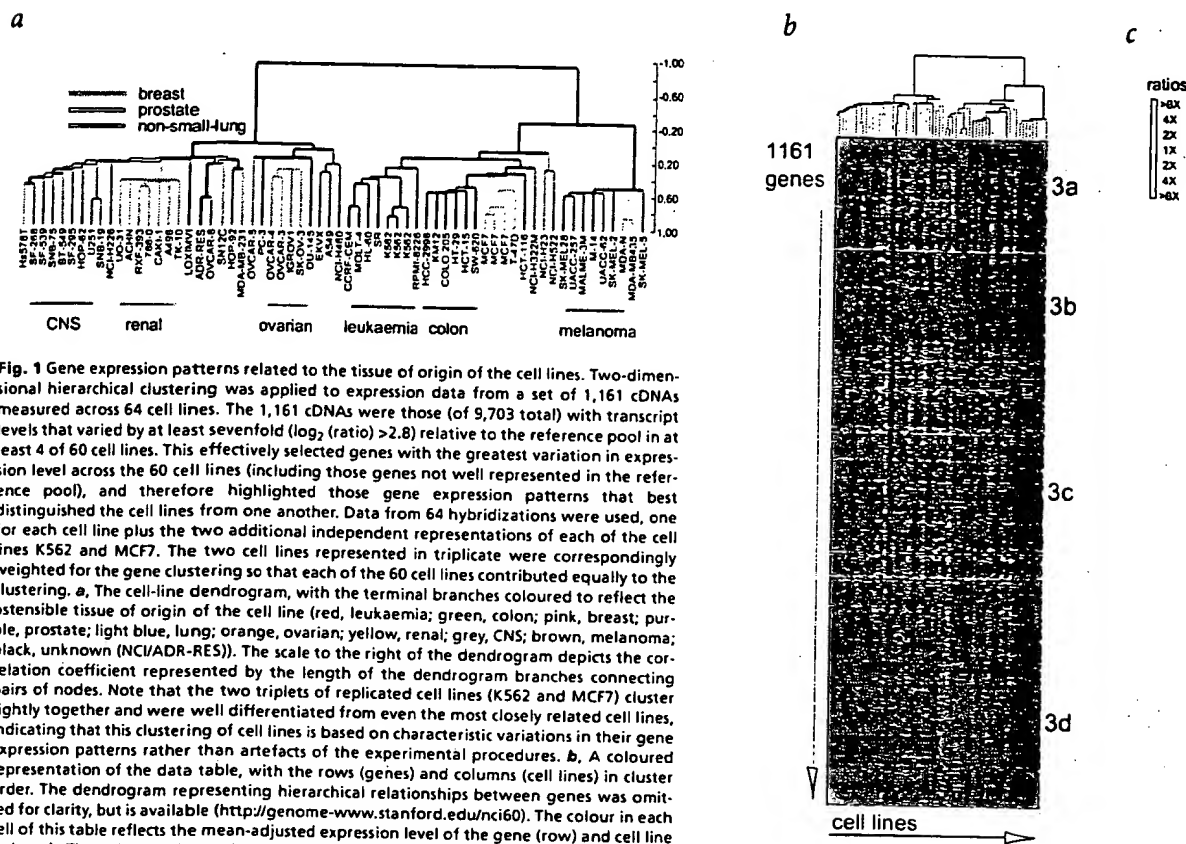


Fig. 1 Gene expression patterns related to the tissue of origin of the cell lines. Two-dimensional hierarchical clustering was applied to expression data from a set of 1,161 cDNAs measured across 64 cell lines. The 1,161 cDNAs were those (of 9,703 total) with transcript levels that varied by at least sevenfold ($\log_2(\text{ratio}) > 2.8$) relative to the reference pool in at least 4 of 60 cell lines. This effectively selected genes with the greatest variation in expression level across the 60 cell lines (including those genes not well represented in the reference pool), and therefore highlighted those gene expression patterns that best distinguished the cell lines from one another. Data from 64 hybridizations were used, one for each cell line plus the two additional independent representations of each of the cell lines K562 and MCF7. The two cell lines represented in triplicate were correspondingly weighted for the gene clustering so that each of the 60 cell lines contributed equally to the clustering. **a**, The cell-line dendrogram, with the terminal branches coloured to reflect the ostensible tissue of origin of the cell line (red, leukaemia; green, colon; pink, breast; purple, prostate; light blue, lung; orange, ovarian; yellow, renal; grey, CNS; brown, melanoma; black, unknown (NCIADR-RES)). The scale to the right of the dendrogram depicts the correlation coefficient represented by the length of the dendrogram branches connecting pairs of nodes. Note that the two triplets of replicated cell lines (K562 and MCF7) cluster tightly together and were well differentiated from even the most closely related cell lines, indicating that this clustering of cell lines is based on characteristic variations in their gene expression patterns rather than artefacts of the experimental procedures. **b**, A coloured representation of the data table, with the rows (genes) and columns (cell lines) in cluster order. The dendrogram representing hierarchical relationships between genes was omitted for clarity, but is available (<http://genome-www.stanford.edu/nci60>). The colour in each cell of this table reflects the mean-adjusted expression level of the gene (row) and cell line (column). The colour scale used to represent the expression ratios is shown. The labels '3a–3d' in (b) refer to the clusters of genes shown in detail in Fig. 3.

from these experiments have been sequence-verified, including all of those referred to here by name.

Each hybridization compared Cy5-labelled cDNA reverse transcribed from mRNA isolated from one of the cell lines with Cy3-labelled cDNA reverse transcribed from a reference mRNA sample. This reference sample, used in all hybridizations, was prepared by combining an equal mixture of mRNA from 12 of the cell lines (chosen to maximize diversity in gene expression as determined primarily from two-dimensional gel studies²). By comparing cDNA from each cell line with a common reference, variation in gene expression across the 60 cell lines could be inferred from the observed variation in the normalized Cy5/Cy3 ratios across the hybridizations.

To assess the contribution of artefactual sources of variation in the experimentally measured expression patterns, K562 and MCF7 cell lines were each grown in three independent cultures, and the entire process was carried out independently on mRNA extracted from each culture. The variance in the triplicate fluorescence ratio measurements approached a minimum when the fluorescence signal was greater than approximately 0.4% of the measurable total signal dynamic range above background in either channel of the hybridization. We selected the subset of spots for which significant signal was present in both the numerator and denominator of the ratios by this criterion to identify the best-measured spots. The pair-wise correlation coefficients for the triplicates of the set of genes that passed this quality control level (6,992 spots included for the MCF7 samples and 6,161 spots for K562) ranged from 0.83 to 0.92 (for graphs and details, see <http://genome-www.stanford.edu/nci60>).

To make the orderly features in the data more apparent, we used a hierarchical clustering algorithm^{19,20} and a pseudo-colour visu-

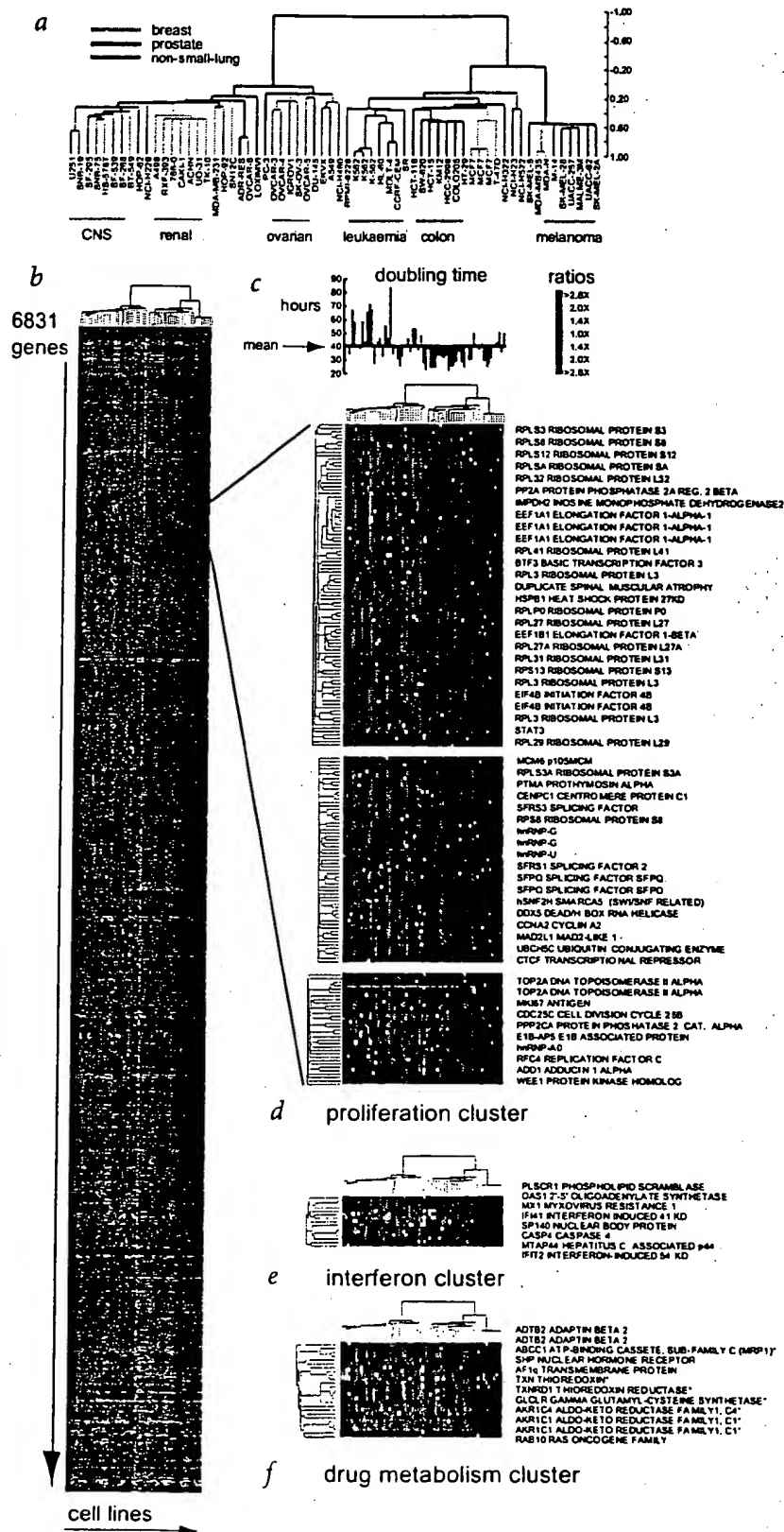
alization matrix^{3,21}. The object of the clustering was to group cell lines with similar repertoires of expressed genes and to group genes whose expression level varied among the 60 cell lines in a similar manner. Clustering was performed twice using different subsets of genes to assess the robustness of the analysis. In one case (Fig. 1), we concentrated on those genes that showed the most variation in expression among the 60 cell lines (1,167 total). A second analysis (Fig. 2) included all spots that were thought to be well measured in the reference set (6,831 spots).

Gene expression patterns related to the histologic origins of the cell lines

The most notable property of the clustered data was that cell lines with common presumptive tissues of origin grouped together (Figs 1a and 2). Cell lines derived from leukaemia, melanoma, central nervous system, colon, renal and ovarian tissue were clustered into independent terminal branches specific to their respective organ types with few exceptions. Cell lines derived from non-small lung carcinoma and breast tumours were distributed in multiple different terminal branches suggesting that their gene expression patterns were more heterogeneous.

Many of these coherent cell line clusters were distinguished by the specific expression of characteristic groups of genes (Fig. 3a–d). For example, a cluster of approximately 90 genes was highly expressed in the melanoma-derived lines (Fig. 3c). This set was enriched for genes with known roles in melanocyte biology, including tyrosinase and dopachrome tautomerase (TYR and DCT; two subunits of an enzyme complex involved in melanin synthesis²²), MART1 (MLANA; which is being investigated as a target for immunotherapy of melanoma²³) and S100- β (S100B; which has been used as an antigenic marker in the diagnosis of

Fig. 2 Gene expression patterns related to other cell-line phenotypes. **a**, We applied two-dimensional hierarchical clustering to expression data from a set of 6,831 cDNAs measured across the 64 cell lines. The 6,831 cDNAs were those with a minimum fluorescence signal intensity of approximately 0.4% of the dynamic range above background in the reference channel in each of the six hybridizations used to establish reproducibility. This effectively selected those spots that provided the most reliable ratio measurements and therefore identified a subset of genes useful for exploring patterns comprised of those whose variation in expression across the 60 cell lines was of moderate magnitude. **b**, Cluster-ordered data table. **c**, Doubling time of cell lines. Cell lines are given in cluster order. Values are plotted relative to the mean. Doubling times greater than the mean are shown in green, those with doubling time less than the mean are shown in red. **d**, Three related gene clusters that were enriched for genes whose expression level variation was correlated with cell line proliferation rate. Each of the three gene clusters (clustered solely on the basis of their expression patterns) showed enrichment for sets of genes involved in distinct functional categories (for example, ribosomal genes versus genes involved in pre-RNA splicing). **e**, Gene cluster in which all characterized and sequence-verified cDNAs encode genes known to be regulated by interferons. **f**, Gene cluster enriched for genes that have been implicated in drug metabolism (indicated by asterisks). A further property of the gene clustering evident here and in Fig. 2 is the strong tendency for redundant representations of the same gene to cluster immediately adjacent to one another, even within larger groups of genes with very similar expression patterns. In addition to illustrating the reproducibility and consistency of the measurements, and providing independent confirmation of many of our measurements, this property also demonstrates that these, and probably all, genes have nearly unique patterns of variation across the 60 cell lines. If this were not the case, and multiple genes had identical patterns of variation, we would not expect to be able to distinguish, by clustering on the basis of expression variation, duplicate copies of individual genes from the other genes with identical expression patterns.



melanoma). LOXIMVI, the seventh line designated as melanoma in the NCI60, did not show this characteristic pattern. Although isolated from a patient with melanoma, LOXIMVI has previously been noted to lack melanin and other markers useful for identification of melanoma cells¹.

Paradoxically, two related cell lines (MDA-MB435 and MDA-N), which were derived from a single patient with breast cancer and have been conventionally regarded as breast cancer cell lines, shared expression of the genes associated with melanoma. MDA-MB435 was isolated from a pleural effusion in a patient with metastatic ductal adenocarcinoma of the breast^{24,25}. It remains possible that the origin of the cell line was a breast cancer, and that its gene expression pattern is related to the neuroendocrine features of some breast cancers²⁶. But our results suggest that this cell line may have originated from a melanoma, raising the possibility that the patient had a co-existing occult melanoma.

The higher-level organization of the cell-line tree—in which groups span cell lines from different tissue types—also reflected shared biological properties of the tissues from which the cell lines were derived. The carcinoma-derived cell lines were divided into major branches that separated those that expressed genes characteristic of epithelial cells from those that expressed genes more typical of stromal cells. A cluster of genes is shown (Fig. 3b) that is most strongly expressed in cell lines derived from colon carcinomas, six of seven ovarian-derived cell lines and the two breast cancer lines positive for the oestrogen receptor. The named genes in this cluster have been implicated in several aspects of epithelial cell biology²⁷. The cluster was enriched for genes whose products are known to localize to the basolateral membrane of epithelial cells, including those encoding components of adherens complexes (for example, desmoplakin (DSP), periplakin (PPL) and plakoglobin (JUP)), an epithelial-expressed cell-cell adhesion molecule (M4S1) and a sodium/hydrogen ion exchanger^{28–31} (SLC9A1). It also contained genes that encode putative transcriptional regulators of epithelial morphogenesis, a human homologue of a *Drosophila melanogaster* epithelial-expressed tumour suppressor (LLGL1) and a homeobox gene thought to control calcium-mediated adherence in epithelial cells^{32,33} (MSX2).

In contrast, a separate, major branch of the cell-line dendrogram (Fig. 1a) included all glioblastoma-derived cell lines, all renal-cell-carcinoma-derived cell lines and the remaining carcinoma-derived lines. The characteristic set of genes expressed in this cluster included many whose products are involved in stromal cell functions (Fig. 3d). Indeed, the two cell lines originally described as 'sarcoma-like' in appearance (Hs578T, breast carcinosarcoma, and SF539, gliosarcoma) expressed most of these genes^{34,35}. Although no single gene was uniformly characteristic of this cluster, each cell line showed a distinctive pattern of expression of genes encoding proteins with roles in synthesis or modification of the extracellular matrix (for example, caldesmon (CALD1), cathepsins, thrombospondin (THBS), lysyl oxidase (LOX) and collagen subtypes). Although the ovarian and most non-small-cell-lung-derived carcinomas expressed genes characteristic of both epithelial cells and stromal cells, they probably clustered with the CNS and renal cell carcinomas in this analysis because genes characteristically expressed in stromal cells were more abundantly represented in this gene set.

Physiological variation reflected in gene expression patterns

A cluster diagram of 6,831 genes (Fig. 2) is useful for exploring clusters of genes whose variation in mRNA levels was not obviously attributable to cell or tissue type. We identified some gene clusters that were enriched for genes involved in specific cellular

processes; the variation in their expression levels may reflect corresponding differences in activity of these processes in the cell lines. For example, a cluster of 1,159 genes (Fig. 2a) included many whose products are necessary for progression through the cell cycle (such as CCNA1, MCM106 and MAD2L1), RNA processing and translation machinery (such as RNA helicases, hnRNPs and translation elongation factors) and traditional pathologic markers used to identify proliferating cells (MKI67). Within this large cluster were smaller clusters enriched for genes with more specialized roles. One cluster was highly enriched for numerous ribosomal genes, whereas another was more enriched for genes encoding RNA-splicing factors. The variation in expression of these ribosomal genes was significantly correlated with variation in the cell doubling time (correlation coefficient of 0.54), supporting the notion that the genes in this cluster were regulated in relation to cell proliferation rate or growth rate in these cell lines.

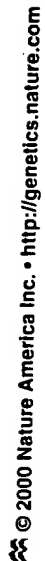
In a smaller gene cluster (Fig. 2d), all of the named genes were previously known to be regulated by interferons^{13,36}. Additional groups of interferon-regulated genes showed distinct patterns of expression (data not shown), suggesting that the NCI60 cell lines exhibited variation in activity of interferon-response pathways, which was reflected in gene expression patterns³⁶.

Another cluster (Fig. 2e) contained several genes encoding proteins with possible interrelated roles in drug metabolism, including glutamate-cysteine ligase (GLCLC, the enzyme responsible for the rate limiting step of glutathione synthesis), thioredoxin (TXN) and thioredoxin reductase (TXNRD1; enzymes involved in regulating redox state in cells), and MRP1 (a drug transporter known to efficiently transport glutathione-conjugated compounds³⁷). The elevated expression of this set of genes in a subset of these cell lines may reflect selection for resistance to chemotherapeutics.

Cell lines facilitate interpretation of gene expression patterns in complex clinical samples

Like many other types of cancer, tumours of the breast typically have a complex histological organization, with connective tissue and leukocytic infiltrates interwoven with tumour cells. To explore the possibility that variation in gene expression in the tumour cell lines might provide a framework for interpreting the expression patterns in tumour specimens, we compared RNA isolated from two breast cancer biopsy samples, a sample of normal breast tissue and the NCI60 cell lines derived from breast cancers (excluding MDA-MB-435 and MDA-N) and leukaemias (Fig. 4). This clustering highlighted features of the gene expression pattern shared between the cancer specimens and individual cell lines derived from breast cancers and leukaemias.

The genes encoding keratin 8 (KRT8) and keratin 19 (KRT19), as well as most of the other 'epithelial' genes defined in the complete NCI60 cell line cluster, were expressed in both of the biopsy samples and the two breast-derived cell lines, MCF-7 and T47D, expressing the oestrogen receptor, suggesting that these transcripts originated in tumour cells with features similar to those of luminal epithelial cells (Fig. 5a). Expression of a set of genes characteristic of stromal cells, including collagen genes (COL3A1, COL5A1 and COL6A1) and smooth muscle cell markers (TAGLN), was a feature shared by the tumour sample and the stromal-like cell lines Hs578T and BT549 (Fig. 5b). This feature of the expression pattern seen in the tumour samples is likely to be due to the stromal component of the tumour. The tumours also shared expression of a set of genes (Fig. 5c) with the multiple myeloma cell line (RPMI-8226), notably including immunoglobulin genes, consistent with the presence of B cells in the tumour (this was confirmed by staining with anti-



nature genetics • volume 24 • march 2000

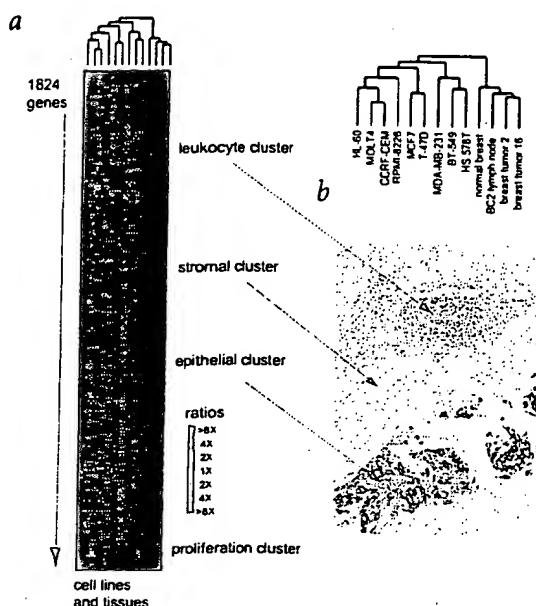


Fig. 4 Comparison of the gene expression patterns in clinical breast cancer specimens and cultured breast cancer and leukaemia cell lines. **a**, Two-dimensional hierarchical clustering applied to gene expression data for two breast cancer specimens, a lymph node metastasis from one patient, normal breast and the NCI60 breast and leukaemia-derived cell lines. The gene expression data from tissue specimens was clustered along with expression data from a subset of the NCI60 cell lines to explore whether features of expression patterns observed in specific lines could be identified in the tissue samples. Labels indicate gene clusters (shown in detail in Fig. 5) that may be related to specific cellular components of the tumour specimens. **b**, Breast cancer specimen 16 stained with anti-keratin antibodies, showing the complex mix of cell types characteristically found in breast tumours. The arrows highlight the different cellular components of this tissue specimen that were distinguished by the gene expression cluster analysis (Fig. 5).

immunoglobulin antibodies; data not shown). Therefore, distinct sets of genes with co-varying expression among the samples (Fig. 4, arrow) appear to represent distinct cell types that can be distinguished in breast cancer tissue. A fourth cluster of genes, more highly expressed in all of the cell lines than in any of the clinical specimens, was enriched for genes present in the 'proliferation' cluster described above (Fig. 5d). The variation in expression of these genes likely paralleled the difference in proliferation rate between the rapidly cycling cultured cell lines and the much more slowly dividing cells in tissues.

Discussion

Newly available genomics tools allowed us to explore variation in gene expression on a genomic scale in 60 cell lines derived from diverse tumour tissues. We used a simple cluster analysis to identify the prominent features in the gene expression patterns that appeared to reflect 'molecular signatures' of the tissue from which the cells originated. The histological characteristics of the cell lines that dominated the clustering were pervasive enough that similar relationships were revealed when alternative subsets of genes were selected for analysis. Additional features of the expression pattern may be related to variation in physiological attributes such as proliferation rate and activity of interferon-response pathways.

The properties of the tumour-derived cell lines in this study have presumably all been shaped by selection for resistance to host defences and chemotherapeutics and for rapid proliferation in the tissue culture environment of synthetic growth media, fetal bovine serum and a polystyrene substratum. But the primary identifiable factor accounting for variation in gene expression patterns among these 60 cell lines was the identity of the tissue from which each cell line was ostensibly derived. For most of the cell lines we examined, neither physiological nor experimental adaptation for growth in culture was sufficient to overwrite the gene expression programs established during differentiation *in vivo*. Nevertheless, the prominence of mesenchymal features in the cell lines isolated from glioblastomas and carcinomas may reflect a selection for the relative ease of establishment of cell lines expressing stromal characteristics, perhaps combined with physiological adaptation to tissue culture conditions^{38–40}.

Biological themes linking genes with related expression patterns may be inferred in many cases from the shared attributes of known genes within the clusters. Uncharacterized cDNAs are likely to encode proteins that have roles similar to those of the known gene products with which they appear to be co-regulated. Still, for several clusters of genes, we were unable to discern a common theme linking the identified members of the cluster. Further exploration of their variation in expression under more diverse conditions and more comprehensive investigation of the physiology of the NCI60 cells may provide insight¹⁰. The relationship of the gene expression patterns to the drug sensitivity patterns measured by the DTP is an example of linking variation in gene expression with more subtle and diverse phenotypic variation¹¹.

The patterns of gene expression measured in the NCI60 cell lines provide a framework that helps to distinguish the cells that express specific sets of genes in the histologically complex breast cancer specimens⁴¹. Although it is now feasible to analyse gene expression in micro-dissected tumour specimens^{42,43}, this observation suggests that it will be possible to explore and interpret some of the biology of clinical tumour samples by sampling them intact. As is useful in conventional morphological pathology, one might be able to observe interactions between a tumour and its microenvironment in this way. These relationships will be clarified by suitable analysis of gene expression patterns from intact as well as dissected tumours^{12,14,15,41}.

Methods

cDNA clones. We obtained the 9,703 human cDNA clones (Research Genetics) used in these experiments as bacterial colonies in 96-well microtitre plates⁹. Approximately 8,000 distinct Unigene clusters (representing nominally unique genes) were represented in this set of clones. All genes identified here by name represent clones whose identities were confirmed by re-sequencing, or by the criteria that two or more independent cDNA clones ostensibly representing the same gene had nearly identical gene expression patterns. A single-pass 3' sequence re-verification was attempted for every clone after re-streaking for single colonies. For a subset of genes for which quality 3' sequence was not obtained, we attempted to confirm identities by 5' sequencing. Of the subset of clones selected for 5' sequence verification on the basis of an interesting pattern of expression (888 total), 331 were correctly identified, 57, incorrectly identified, and 500, indeterminate (poor quality sequence). We estimated that 15%–20% of array elements contained DNA representing more than one clone per well. So far, the identities of ~3,000 clones have been verified. The full list of clones used and their nominal identities are available (gene names preceded by the designation "SID#" (Stanford Identification) represent clones whose identities have not yet been verified; <http://genome-www.stanford.edu:8000/nci60>).

Production of cDNA microarrays. The arrays used in this experiment were produced at Synteni Inc. (now Incyte Pharmaceuticals). Each insert was amplified from a bacterial colony by sampling 1 µl of bacterial media and performing PCR amplification of the insert using consensus primers for the three plasmids represented in the clone set (5'-TTGTAACGACG GCCAGTG-3', 5'-CACACAGGAACAGCTATG-3'). Each PCR product

(100 µl) was purified by gel exclusion, concentrated and resuspended in 3×SSC (10 µl). The PCR products were then printed on treated glass microscope slides using a robot with four printing tips. Detailed protocols for assembling and operating a microarray printer, and printing and experimental application of DNA microarrays are available (<http://cmgm.stanford.edu/pbrown>).

Preparation of mRNA and reference pool. Cell lines were grown from NCI DTP frozen stocks in RPMI-1640 supplemented with phenol red, glutamine (2 mM) and 5% fetal calf serum. To minimize the contribution of variations in culture conditions or cell density to differential gene expression, we grew each cell line to 80% confluence and isolated mRNA 24 h after transfer to fresh medium. The time between removal from the incubator and lysis of the cells in RNA stabilization buffer was minimized (<1 min). Cells were lysed in buffer containing guanidium isothiocyanate and total RNA was purified with the RNeasy purification kit (Qiagen). We purified mRNA as needed

using a poly(A) purification kit (Oligotex, Qiagen) according to the manufacturer's instructions. Denaturing agarose gel electrophoresis assessed the integrity and relative contamination of mRNA with ribosomal RNA.

The breast tumours were surgically excised from patients and rapidly transported to the pathology laboratory, where samples for microarray analysis were quickly frozen in liquid nitrogen and stored at -80 °C until use. A frozen tumour specimen was removed from the freezer, cut into small pieces (~50–100 mg each), immediately placed into 10–12 ml of Trizol reagent (Gibco-BRL) and homogenized using a PowerGen 125 Tissue Homogenizer (Fisher Scientific), starting at 5,000 r.p.m. and gradually increasing to ~20,000 r.p.m. over a period of 30–60 s. We processed the Trizol/tumour homogenate as described in the Trizol protocol, including an initial step to remove fat. Once total RNA was obtained, we isolated mRNA with a FastTrack 2.0 kit (Invitrogen) using the manufacturer's protocol for isolating mRNA starting from total RNA. The normal breast samples were obtained from Clontech.

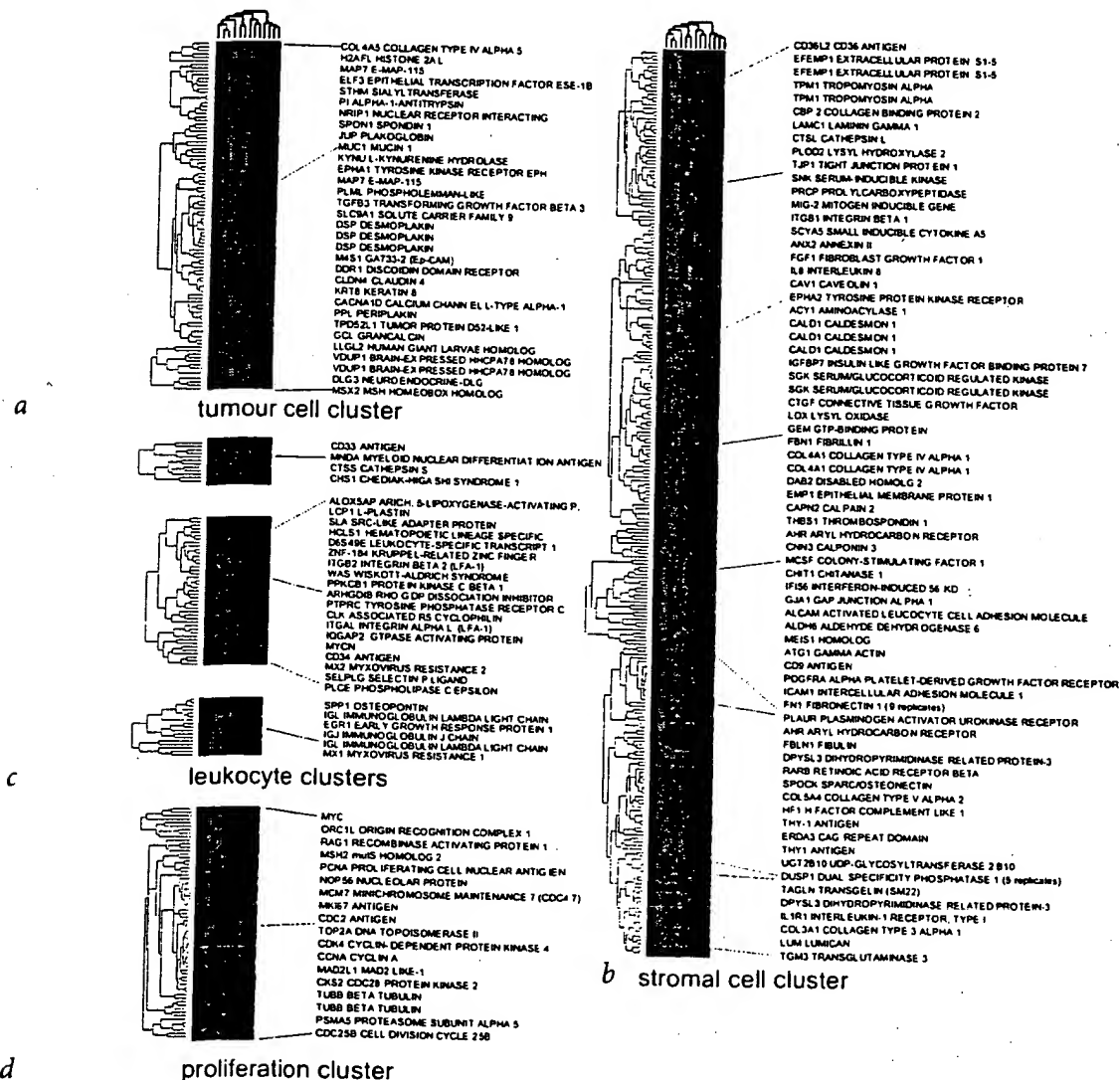


Fig. 5 Histologic features of breast cancer biopsies can be recognized and parsed based on gene expression patterns. Enlargements of the regions of the cluster diagram in Fig. 4 showing gene clusters enriched for genes expressed in different cell types in the breast cancer specimens, as distinguished by clustering with the cultured cell lines. **a**, A cluster including many genes characteristic of epithelial cells expressed in cell lines (T47D and MCF7) derived from breast cancer positive for the oestrogen receptor and tumours. **b**, Genes expressed in cell lines derived from breast cancer with stromal cell characteristics (Hs578T and BT549) and tumour specimens. Expression of these genes in the tumour samples may reflect the presence of myofibroblasts in the cancer specimen stroma. **c**, Genes expressed in leukocyte-derived cell lines, showing common leukocyte, and separate 'myeloid' and 'B-cell', gene clusters. **d**, Genes that were relatively highly expressed in all cell lines compared with the tumour specimens and normal breast. The higher expression of this set of genes involved in cell cycle transit in the cell lines is likely to reflect the higher proliferative rate of cells cultured in the presence of serum compared with the average proliferation rate of cells in the biopsied tissue.

We combined mRNA from the following cells in equal quantities to make the reference pool: HL-60 (acute myeloid leukaemia) and K562 (chronic myeloid leukaemia); NCI-H226 (non-small-cell-lung); COLO 205 (colon); SNB-19 (central nervous system); LOX-IMVI (melanoma); OVCAR-3 and OVCAR-4 (ovarian); CAKI-1 (renal); PC-3 (prostate); and MCF7 and Hs578T (breast). The criterion for selection of the cell lines in the reference are described in detail in the accompanying manuscript¹².

Doubling-time calculations. We calculated doubling times based on routine NCI60 cell line compound screening data; and they reflect the doubling times for cells inoculated into 96-well plates at the screening inoculation densities and grown in RPMI 1640 medium supplemented with 5% fetal bovine serum for 48 h. We measured cell populations using sulforhodamine B optical density measurement assay. The doubling time constant k was calculated using the equation: $N/N_0 = e^{kt}$, where N_0 is optical density for control (untreated) cells at time zero, N is optical density for control cells after 48-h incubation, and t is 48 h. The same equation was then used with the derived k to calculate the doubling time t by setting $N/N_0 = 2$. For a given cell line, we obtained N_0 and N values by averaging optical densities ($N > 6,000$) obtained for each cell line for a year's screening. Data and experimental details are available (<http://dtp.nci.nih.gov>).

Preparation and hybridization of fluorescent labelled cDNA. For each comparative array hybridization, labelled cDNA was synthesized by reverse transcription from test cell mRNA in the presence of Cy5-dUTP, and from the reference mRNA with Cy3-dUTP, using the Superscript II reverse-transcription kit (Gibco-BRL). For each reverse transcription reaction, mRNA (2 µg) was mixed with an anchored oligo-dT (d-20T-d(AGC)) primer (4 µg) in a total volume of 15 µl, heated to 70 °C for 10 min and cooled on ice. To this sample, we added an unlabelled nucleotide pool (0.6 µl; 25 mM each dATP, dCTP, dGTP, and 15 mM dTTP), either Cy3 or Cy5 conjugated dUTP (3 µl; 1 mM; Amersham), 5×first-strand buffer (6 µl; 250 mM Tris-HCl, pH 8.3, 375 mM KCl, 15 mM MgCl₂), 0.1 M DTT (3 µl) and 2 µl of Superscript II reverse transcriptase (200 µl). After a 2-h incubation at 42 °C, the RNA was degraded by adding 1 N NaOH (1.5 µl) and incubating at 70 °C for 10 min. The mixture was neutralized by adding of 1 N HCl (1.5 µl), and the volume brought to 500 µl with TE (10 mM Tris, 1 mM EDTA). We added Cot1 human DNA (20 µg; Gibco-BRL), and purified the probe by centrifugation in a Centricon-30 micro-concentrator (Amicon). The two separate probes were combined, brought to a volume of 500 µl, and concentrated again to a volume of less than 7 µl. We added 10 µg/µl poly(A) RNA (1 µl; Sigma) and tRNA (10 µg/µl; Gibco-BRL) were added, and adjusted the volume to 9.5 µl with distilled water. For final probe preparation, 20×SSC (2.1 µl; 1.5 M NaCl, 150 mM NaCitrate, pH 8.0) and 10% SDS (0.35 µl) were added to a total final volume of 12 µl. The probes were denatured by heating for 2 min at 100 °C, incubated at 37 °C for 20–30 min, and placed on the array under a 22 mm×22 mm glass coverslip. We incubated slides overnight at 65 °C for 14–18 h in a custom slide chamber with humidity maintained by a small reservoir of 3×SSC. Arrays were washed by submersion and agitation for 2–5 min in 2×SSC with 0.1% SDS, followed by 1×SSC and then 0.1×SSC. The arrays were "spun dry" by centrifugation for 2 min in a slide-rack in a Beckman GS-6 tabletop centrifuge in Microplus carriers at 650 r.p.m. for 2 min.

Array quantitation and data processing. Following hybridization, arrays were scanned using a laser-scanning microscope (ref. 17; <http://cmgm.stanford.edu/pbrown>). Separate images were acquired for Cy3 and Cy5. We carried out data reduction with the program ScanAlyze (M.B.E., available

at <http://rana.stanford.edu/software>). Each spot was defined by manual positioning of a grid of circles over the array image. For each fluorescent image, the average pixel intensity within each circle was determined, and a local background was computed for each spot equal to the median pixel intensity in a square of 40 pixels in width and height centred on the spot centre, excluding all pixels within any defined spots. Net signal was determined by subtraction of this local background from the average intensity for each spot. Spots deemed unsuitable for accurate quantitation because of array artefacts were manually flagged and excluded from further analysis. Data files generated by ScanAlyze were entered into a custom database that maintains web-accessible files. Signal intensities between the two fluorescent images were normalized by applying a uniform scale factor to all intensities measured for the Cy5 channel. The normalization factor was chosen so that the mean $\log(Cy3/Cy5)$ for a subset of spots that achieved a minimum quality parameter (approximately 6,000 spots) was 0. This effectively defined the signal-intensity-weighted 'average' spot on each array to have a Cy3/Cy5 ratio of 1.0.

Cluster analysis. We extracted tables (rows of genes, columns of individual microarray hybridizations) of normalized fluorescence ratios from the database. Various selection criteria, discussed in relation to each data set, were applied to select subsets of genes from the 9,703 cDNA elements on the arrays. Before clustering and display, the logarithm of the measured fluorescence ratios for each gene were centred by subtracting the arithmetic mean of all ratios measured for that gene. The centring makes all subsequent analyses independent of the amount of each gene's mRNA in the reference pool.

We applied a hierarchical clustering algorithm separately to the cell lines and genes using the Pearson correlation coefficient as the measure of similarity and average linkage clustering^{3,19–21}. The results of this process are two dendrograms (trees), one for the cell lines and one for the genes, in which very similar elements are connected by short branches, and longer branches join elements with diminishing degrees of similarity. For visual display the rows and columns in the initial data table were reordered to conform to the structures of the dendrograms obtained from the cluster analysis. Each cell in the cluster-ordered data table was replaced by a graded colour (pure red through black to pure green), representing the mean-adjusted ratio value in the cell. Gene labels in cluster diagrams are displayed here only for genes that were represented in the microarray by sequence-verified cDNAs. A complete software implementation of this process is available (<http://rana.stanford.edu/software>), as well as all clustering results (<http://genome-www.stanford.edu/nci60>).

Acknowledgements

We thank members of the Brown and Botstein labs for helpful discussions. This work was supported by the Howard Hughes Medical Institute and a grant from the National Cancer Institute (CA 077097). The work of U.S. and J.N.W. was supported in part by a grant from the National Cancer Institute Breast Cancer Think Tank. D.T.R. is a Walter and Idun Berry Fellow. M.B.E. is an Alfred P. Sloan Foundation Fellow in Computational Molecular Biology. C.M.P. is a SmithKline Beecham Pharmaceuticals Fellow of the Life Science Research Foundation. P.O.B. is an Associate Investigator of the Howard Hughes Medical Institute.

Received 20 July 1999; accepted 13 January 2000.

1. Stinson, S.F. et al. Morphological and immunocytochemical characteristics of human tumor cell lines for use in a disease-oriented anticancer drug screen. *Anticancer Res.* **12**, 1035–1053 (1992).
2. Myers, T.G. et al. A protein expression database for the molecular pharmacology of cancer. *Electrophoresis* **18**, 647–653 (1997).
3. Weinstein, J.N. et al. An information-intensive approach to the molecular pharmacology of cancer. *Science* **275**, 343–349 (1997).
4. Monks, A., Scudiero, D.A., Johnson, G.S., Paull, K.D. & Sausville, E.A. The NCI anticancer drug screen: a smart screen to identify effectors of novel targets. *Anticancer Drug Des.* **12**, 533–541 (1997).
5. Paull, K.D. et al. Display and analysis of patterns of differential activity of drugs against human tumor cell lines: development of mean graph and COMPARE algorithm. *J. Natl Cancer Inst.* **81**, 1088–1092 (1989).
6. Weinstein, J.N. et al. Neural computing in cancer drug development: predicting mechanism of action. *Science* **258**, 447–451 (1992).
7. van Osdel, W.W., Myers, T.G., Paull, K.D., Kohn, K.W. & Weinstein, J.N. Use of the Kohonen self-organizing map to study the mechanisms of action of chemotherapeutic agents. *J. Natl Cancer Inst.* **86**, 1853–1859 (1994).
8. DeRisi, J.L., Iyer, V.R. & Brown, P.O. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**, 680–686 (1997).
9. Iyer, V.R. et al. The transcriptional program in the response of human fibroblasts to serum. *Science* **283**, 83–87 (1999).
10. Brown, P.O. & Botstein, D. Exploring the new world of the genome with DNA microarrays. *Nature Genet.* **21** (suppl.), 33–37 (1999).
11. Scherf, U. et al. A gene expression database for the molecular pharmacology of cancer. *Nature Genet.* **24**, 236–244 (2000).
12. Khan, J. et al. Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays. *Cancer Res.* **58**, 5009–5013 (1998).
13. Der, S.D., Zhou, A., Williams, B.R. & Silverman, R.H. Identification of genes differentially regulated by interferon- α , - β or - γ using oligonucleotide arrays. *Proc. Natl Acad. Sci. USA* **95**, 15623–15628 (1998).
14. Alon, U. et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl Acad. Sci. USA* **96**, 6745–6750 (1999).
15. Wang, K. et al. Monitoring gene expression profile changes in ovarian carcinomas using cDNA microarray. *Gene* **229**, 101–108 (1999).
16. Tamayo, P. et al. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl Acad. Sci. USA* **96**, 2907–2912 (1999).
17. Shalon, D., Smith, S.J. & Brown, P.O. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res.* **6**, 639–645 (1996).
18. Eisen, M.B. & Brown, P.O. DNA arrays for analysis of gene expression. *Methods Enzymol.* **303**, 179–205 (1999).
19. Sokal, R.R. & Sneath, P.H.A. *Principles of Numerical Taxonomy* (W.H. Freeman, San Francisco, 1963).
20. Hartigan, J.A. *Clustering Algorithms* (Wiley, New York, 1975).
21. Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA* **95**, 14863–14868 (1998).
22. del Marmol, V. & Beermann, F. Tyrosinase and related proteins in mammalian pigmentation. *FEBS Lett.* **381**, 165–168 (1996).
23. Kawakami, Y. et al. The use of melanosomal proteins in the immunotherapy of melanoma. *J. Immunother.* **21**, 237–246 (1998).
24. Cailleau, R., Olive, M. & Cruciger, Q.V. Long-term human breast carcinoma cell lines of metastatic origin: preliminary characterization. *In Vitro* **14**, 911–915 (1978).
25. Brinkley, B.R. et al. Variations in cell form and cytoskeleton in human breast carcinoma cells in vitro. *Cancer Res.* **40**, 3118–3129 (1980).
26. Nesland, J.M., Holm, R., Johannessen, J.V. & Gould, V.E. Neuroendocrine differentiation in breast lesions. *Pathol. Res. Pract.* **183**, 214–221 (1988).
27. Davies, J.A. & Garrod, D.R. Molecular aspects of the epithelial phenotype. *Bioessays* **19**, 699–704 (1997).
28. Garrod, D., Chidgey, M. & North, A. Desmosomes: differentiation, development, dynamics and disease. *Curr. Opin. Cell Biol.* **8**, 670–678 (1996).
29. Cowin, P. & Burke, B. Cytoskeleton-membrane interactions. *Curr. Opin. Cell Biol.* **8**, 56–65 (1996); erratum: **8**, 244 (1996).
30. Litvinov, S.V. et al. Epithelial cell adhesion molecule (Ep-CAM) modulates cell-cell interactions mediated by classic cadherins. *J. Cell Biol.* **139**, 1337–1348 (1997).
31. Helmle-Kolb, C. et al. Na/H exchange activities in NHE1-transfected OK-cells: cell polarity and regulation. *Plügers Arch.* **425**, 34–40 (1993); erratum: **427**, 387 (1994).
32. Manfrulli, P., Arquier, N., Hanratty, W.P. & Semeriva, M. The tumor suppressor gene, *lethal(2)giant larvae* (*l(2)g1*), is required for cell shape change of epithelial cells during *Drosophila* development. *Development* **122**, 2283–2294 (1996).
33. Lincecum, J.M., Fannon, A., Song, K., Wang, Y. & Sassoon, D.A. Msh homeobox genes regulate cadherin-mediated cell adhesion and cell-cell sorting. *J. Cell Biochem.* **70**, 22–28 (1998).
34. Hackett, A.J. et al. Two syngeneic cell lines from human breast tissue: the aneuploid mammary epithelial (Hs578T) and the diploid myoepithelial (Hs578Bst) cell lines. *J. Natl Cancer Inst.* **58**, 1795–1806 (1977).
35. Rutka, J.T. et al. Establishment and characterization of a cell line from a human gliosarcoma. *Cancer Res.* **46**, 5893–5902 (1986).
36. Nguyen, H., Hiscott, J. & Pitha, P.M. The growing family of interferon regulatory factors. *Cytokine Growth Factor Rev.* **8**, 293–312 (1997).
37. Moscow, J.A., Schneider, E., Ivy, S.P. & Cowan, K.H. Multidrug resistance. *Cancer Chemother. Biol. Response Modif.* **17**, 139–177 (1997).
38. Smith, H.S. & Hackett, A.J. The use of cultured human mammary epithelial cells in defining malignant progression. *Ann. N.Y. Acad. Sci.* **464**, 288–300 (1986).
39. Rutka, J.T. et al. Establishment and characterization of five cell lines derived from human malignant gliomas. *Acta Neuropathol.* **75**, 92–103 (1987).
40. Ronnov-Jessen, L., Petersen, O.W. & Bissell, M.J. Cellular changes involved in conversion of normal to malignant breast: importance of the stromal reaction. *Physiol. Rev.* **76**, 69–125 (1996).
41. Perou, C.M. et al. Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc. Natl Acad. Sci. USA* **96**, 9212–9217 (1999).
42. Bonner, R.F. et al. Laser capture microdissection: molecular analysis of tissue. *Science* **278**, 1481–1483 (1997).
43. Sgroi, D.C. et al. In vivo gene expression profile analysis of human breast cancer progression. *Cancer Res.* **59**, 5656–5661 (1999).

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

DECLARATION OF JOHN C. ROCKETT, Ph.D.
UNDER 37 C.F.R. § 1.132

I, JOHN COUGHLIN ROCKETT III, Ph.D., declare and state as follows:

1. Since 1995 I have been engaged full-time in molecular toxicology research, with an emphasis on the application of expression profiling techniques, including but not limited to nucleic acid microarray expression profiling techniques, to studies of the mechanisms of toxicant action and to the design of assays to monitor toxicant exposure.
2. My curriculum vitae, including my list of publications, is attached hereto as Exhibit A.
3. For the past 5 years, my work has focused primarily on analyzing the effects of potentially hazardous environmental agents, such as heat, water disinfectant byproducts, and conazole fungicides on the male reproductive tract. Although we are interested in the basic mechanisms of action of such toxicants, we also have two practical goals in mind: first, to identify individual agents and families of agents that adversely affect male reproductive development and function, and second, to develop methods for monitoring human exposure to such agents, particularly methods capable of identifying toxicant exposure at an early stage.
4. I have relied on expression profiling as a principal approach to these goals. Expression profiling, by

reporting the expression levels of thousands of genes simultaneously, gives us an opportunity to identify and group toxicants based on similarities in the patterns of gene expression they induce in cells and tissues; the gene expression profiles induced by treatment with known testicular toxins serve as standards, molecular signatures or molecular fingerprints as it were, against which the patterns of gene expression induced by agents of unknown toxicity may be compared and judged. In addition, gene expression profiling may give us the opportunity to detect toxicity before more gross phenotypic changes become manifest.

5. In keeping with this research emphasis, I have until recently:

served on the Microarray Technical Subcommittee of the United States Environmental Protection Agency (EPA) Genomics Task Force; and

served on the Scientific Committee for the conference series on "Critical Assessment of Techniques for Microarray Data Analysis," held annually at Duke University, Durham, NC;

and I currently

serve on the Technical Committee on the Application of Genomics to Mechanism-Based Risk Assessment of the International Life Sciences Institute's Health and Environmental Sciences Institute,

serve on the Genomics and Proteomics Committee of the National Health and Environmental Effects Research Laboratory of the EPA's Office of Research and Development,

belong to the [North Carolina Research] Triangle Array Users Group,

belong to the Molecular Biology
Speciality Section of the Society of Toxicology,
and

belong to the Triangle Consortium for
Reproductive Biology.

In addition, I am the principal investigator on a cooperative research and development agreement (CRADA) entitled "Development of a Genetic Test for Male Factor Infertility." Prior to this, I was a co-principal investigator on a materials cooperative research and development agreement (MCRADA) to print oligonucleotide-based microarrays; and from 1999 - 2002, I was coinvestigator on a CRADA to develop gene microarrays for toxicology applications.

6. I presume the reader's familiarity with the basic construction and operation of microarrays. For purposes of the discussion to follow, I use the phrase "nucleic acid microarray" and, equivalently, the term "microarray" to refer generically to the various types of nucleic acid microarray that include immobilized nucleic acid probes of sufficient length to permit specific binding, with minimal cross-hybridization, to the probe's cognate transcript, whether the transcript is in the form of RNA or DNA. Although this definition excludes microarrays having shorter probes, such as the 20-mer probes of arrays manufactured by Affymetrix, Inc., many of the comments that follow nonetheless apply to such microarrays as well.

7. Although my own work with microarrays dates back only to 1998, and high density spotted nucleic acid

microarrays themselves date back perhaps only to 1995,¹ microarrays are by no means the only, nor the first, expression profiling tool. As I describe in detail in my *Xenobiotica* review,² there are a number of other differential expression analysis technologies that precede the development of microarrays, some by decades, and that have been applied to drug metabolism and toxicology research, including:

(1) differential screening; (2) subtractive hybridization, including variants such as chemical cross-linking subtraction, suppression-PCR subtractive hybridization and representational difference analysis; (3) differential display; (4) restriction endonuclease facilitated analyses, including serial analysis of gene expression (SAGE) and gene expression fingerprinting; and (5) EST analysis.

8. In my own earlier research, I used both reverse-transcriptase polymerase chain reaction (RT-PCR) and suppression-PCR subtractive hybridization (SSH) to study patterns of differential gene expression caused by hepatic challenge with nongenotoxic and genotoxic hepatotoxins.³

¹ Schena et al., "Quantitative monitoring of gene expression patterns with a complementary DNA microarray," *Science* 270:467-470 (1995), attached hereto as Exhibit B.

² Rockett et al., "Differential gene expression in drug metabolism and toxicology: practicalities, problems and potential," *Xenobiotica* 29:655-691 (1999) (hereinafter, "*Xenobiotica* review"), attached hereto as Exhibit C.

³ See, e.g., Rockett et al., "Molecular profiling of non-genotoxic carcinogenesis using differential display reverse transcription polymerase chain reaction (ddRT-PCR)," *European J. Drug Metabolism & Pharmacokinetics* 22(4):329-33 (1997), and Rockett et al., "Use of a suppression-PCR subtractive hybridization method to identify gene species which demonstrate altered expression in male rat and guinea pig livers following 3-day exposure to [4-chloro-6-(2,3-xylidino)-2-pyrimidinylthio] acetic acid," *Toxicology* 144(1-3):13-29 (2000), attached hereto respectively as Exhibits D and E.

9. These older transcript expression profiling techniques provide analogous expression data, but with far lower throughput.

10. It has been well-established, at least since the introduction of high density spotted microarrays in 1995, that:

(i) each probe on the microarray, with careful design and sufficient length, and with sufficiently stringent hybridization and wash conditions, binds specifically and with minimal cross-hybridization, to the probe's cognate transcript;

(ii) each additional probe makes an additional transcript newly detectable by the microarray, increasing the detection range, and thus versatility, of this analytical device for gene expression profiling;⁴

(iii) it is not necessary that the biological function be known in order for the gene,

⁴ The compelling logic of this proposition has likely motivated the remarkably rapid progress from the earliest high density spotted arrays in 1995 (Schena et al., "Quantitative monitoring of gene expression patterns with a complementary DNA microarray," *Science* 270:467-470 (1995), attached hereto as Exhibit B), to the first whole genome arrays in 1997 (Lashkari et al., "Yeast microarrays for genome wide parallel genetic and gene expression analysis," *Proc. Natl. Acad. Sci. USA* 94(24):13057-62 (1997) and DeRisi et al., "Exploring the metabolic and genetic control of gene expression on a genomic scale," *Science* 278(5338):680-6 (1997), attached hereto as Exhibits F and G, respectively), to the concurrent announcement by two companies earlier this month of their respective commercial introductions of single chip human whole genome arrays (Pollack, "Human Genome Placed on Chip; Biotech Rivals Put it Up for Sale," *The New York Times*, Thursday, October 2, 2003 (Business Day), attached hereto as Exhibit H; "Agilent Technologies ships whole human genome on single microarray to gene expression customers for evaluation," Press Release, Agilent Technologies, October 2, 2003, attached hereto as Exhibit I; "Affymetrix Announces Commercial Launch of Single Array for Human Genome Expression Analysis; More Than 1 Million Probes Analyze Expression Levels of Nearly 50,000 RNA Transcripts and Variants on a Single Array the Size of a Thumbnail," Press Release, Affymetrix, October 2, 2003, attached hereto as Exhibit J).

or a fragment of the gene, to prove useful as a probe on a microarray to be used for expression analysis;

(iv) failure of a probe to detect changes in expression of its cognate gene does not diminish the usefulness of the probe on the microarray; and

(iv) failure of a probe to detect a particular transcript in any single experiment does not deprive the probe of usefulness to the community of users who would use this research tool.

These principles also apply to transcript expression profiling techniques that antedate the development of high density spotted microarrays, and accordingly were well-understood prior to 1995.

11. Moreover, expression profiling is not limited to the measurement of mRNA transcript levels. It is widely understood among molecular and cellular biologists that protein expression levels provide complementary profiles for any given cell and cellular state. Although I cannot claim credit for having coined the phrase, I have written that the difference between transcript expression profiling and protein expression profiling is that "transcriptomics indicates what *should happen*, and proteomics shows what *is happening*."⁵

12. For decades, such protein expression profiles have been generated using two dimensional polyacrylamide gel

⁵ Rockett, "Macroresults through Microarrays," *Drug Discovery Today* 7:804 - 805 (2002) (emphasis added), attached hereto as Exhibit K.

electrophoresis (2D-PAGE), and used, among other things, to study drug effects.⁶

13. Although the protein expression profiles produced by 2D-PAGE analysis are analogous to the transcript expression profiles provided by nucleic acid microarrays, an even closer analogy is perhaps offered by antibody microarrays; as I note in my *Drug Discovery Today* commentary, such antibody microarrays date back to the work of Roger Ekins in the mid- to late-1980s.⁷

14. The principles in paragraph 10 also apply to protein expression profiling analyses, particularly to analyses performed using antibody microarrays. Thus, as with nucleic acid microarrays, the greater the number of proteins detectable, the greater the power of the technique; the absence or failure of a protein to change in expression levels does not diminish the usefulness of the method; and prior knowledge of the biological function of the protein is not required. As applied to protein expression profiling, these principles have been well understood since at least as early as the 1980s.

15. Both gene and protein expression profiling are particularly useful to the toxicologist, especially in the pharmaceutical industry. Accordingly, I made the following

⁶ See, e.g., Anderson et al., "A two-dimensional gel database of rat liver proteins useful in gene regulation and drug effects studies," *Electrophoresis* 12:907 - 930 (1991), attached hereto as Exhibit L.

⁷ See Ekins et al., *J. Bioluminescence Chemiluminescence* 5:59-78 (1989); Ekins et al., *Clin. Chem.* 37: 1955-1965 (1991); and Ekins, U.S. Patent Nos. 5,432,099, 5,807,755, and 5,837,551, attached hereto respectively as Exhibits M to Q.

statements in my *Xenobiotica* review, written in the summer of 1998:

[I]n the field of chemical-induced toxicity, it is now becoming increasingly obvious that most adverse reactions to drugs and chemicals are the result of multiple gene regulation, some of which are causal and some of which are casually-related to the toxicological phenomenon *per se*. This observation has led to an upsurge in interest in gene-profiling technologies which differentiate between the control and toxin-treated gene pools in target tissues and is, therefore, of value in rationalizing the molecular mechanisms of xenobiotic-induced toxicity.

Knowledge of toxin-dependent gene regulation in target tissues is not solely an academic pursuit as much interest has been generated in the pharmaceutical industry to harness this technology in the early identification of toxic drug candidates, thereby shortening the developmental process and contributing substantially to the safety assessment of new drugs.

For example, if the gene profile in response to say a testicular toxin that has been well-characterized *in vivo* could be determined in the testis, then this profile would be representative of all new drug candidates which act via this specific molecular mechanism of toxicity, thereby providing a useful and coherent approach to the early detection of such toxicants.

Whereas it would be informative to know the identity and functionality of all genes up/down regulated by such toxicants, this would appear a longer term goal, as the majority of human genes have not yet been sequenced, far less their functionality determined. However, the current use of gene profiling yields a pattern of gene changes for a xenobiotic of unknown toxicity which may be matched to that of well-characterized toxins, thus alerting the toxicologist to possible *in vivo* similarities between the unknown and the standard. . . .

* * *

Despite the development of multiple technological advances which have recently brought the field of gene expression profiling to the forefront of molecular analysis, recognition of the importance of differential gene expression and characterization of differentially expressed genes has existed for many years.

16. As noted in the preceding excerpt from my *Xenobiotica* review, expression profiling in toxicology studies yield patterns of changes that are characteristic of an agent of unknown toxicity, which patterns may usefully be matched to those of well-characterized toxins.

17. In the context of such patterns of gene expression, each additional gene-specific probe provides an additional signal that could not otherwise have been detected, giving a more comprehensive, robust, higher resolution -- and thus more useful -- pattern than otherwise would have been possible.⁸

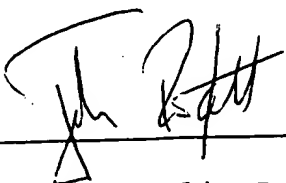
18. It is my opinion, therefore, based on the state of the art in toxicology at least since the mid-1990s -- and as regards protein profiling, even earlier -- that disclosure of the sequence of a new gene or protein, with or without knowledge of its biological function, would have been

⁸ In a sense, each gene-specific probe used in such an analysis is analogous to a different one of the many parts of an engine, with each individual part, or subcombinations of such parts, deriving at least part of their usefulness from the utility of the completed combination, the functioning engine.

sufficient information for a toxicologist to use the gene and/or protein in expression profiling studies in toxicology.

19. The statements made in this declaration represent my individual views and are not intended to represent the opinion of my employer, the United States Environmental Protection Agency, or of any other branch of the federal government. Other than my current engagement to provide this declaration, I have neither had, nor currently have, financial ties to, or financial interest in, Incyte Corporation. I am not myself an inventor on any patent application claiming a gene or gene fragment.

20. I declare further that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true, and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and may jeopardize the validity of any patent application in which this declaration is filed or any patent that issues thereon.



John Coughlin Rockett III, Ph.D.

10-17-03

Date

CURRICULUM VITAE

PERSONAL DETAILS

Name: John Coughlin Rockett III

Nationality: USA

Work Address: United States Environmental Protection Agency
National Health and Environmental Effects Research Laboratory
Reproductive Toxicology Division (MD-72)
Gamete and Early Embryo Biology Branch
Research Triangle Park
NC 27711
USA

Work Telephone: +001 (919) 541 2678

Work Fax: +001 (919) 541 4017

Work E-mail: rockett.john@epa.gov

Employment and Higher Education

CURRENT POSITION (12/00-present)

Research Biologist
Gamete and Early Embryo Biology Branch (MD-72)
Reproductive Toxicology Division
National Health and Environmental Effects Research Laboratory
US Environmental Protection Agency
Research Triangle Park
NC 27711
USA

PREVIOUS POSITIONS

8/98-12/00: NHEERL Post-Doctoral Research Fellow, Gamete and Early Embryo Biology Branch, Reproductive Toxicology Division, National Health and Environmental Effects Research Laboratory, United States Environmental Protection Agency, Research Triangle Park, NC, USA.

Supervisors: Dr Sally P. Darney (Scientific publications under Sally D. Perreault) and Dr David J. Dix.

5/95-7/98: Rhone-Poulenc Post-Doctoral Research Fellow, Molecular Toxicology Group, School of Biological Sciences, University of Surrey, Guildford, Surrey, England.
Supervisor: Prof. G. Gordon Gibson.

EDUCATION

Ph.D., 1995 - University of Warwick, Coventry, W. Midlands, England
Title: Transforming Growth Factor- β and Immune Recognition Molecules in Oesophageal Cancer.
Supervisors: Dr Alan G. Morris (University of Warwick) and Dr S. Jane Darnton (Birmingham Heartlands Hospital)

B.Sc. (Hons.), 1991 - University of Warwick, Coventry, W. Midlands, England.
Degree: Microbiology and Microbial Technology (with intercalated year in industry), Class 2i.
Tutor: Professor Howard Dalton.

PROFESSIONAL ACTIVITIES

Membership of Professional Societies:

Society of Toxicology (Inc. Molecular Biology Speciality Section) (2001-present)
Science Advisory Board (2001-present)
North Carolina Chapter of the Society of Toxicology (1999-present)
Triangle Consortium for Reproductive Biology (1999-present)
Triangle Array Users Group (1999-present)
Institute of Biology (U.K.) (1989 - present)
British Toxicology Society (1996 - 2000)
Biochemical Society (U.K.) (1992-1995)
British Society for Immunology (1992-1995)

Membership of Scientific Committees:

International Life Sciences Institute's (ILSI) Health and Environmental Sciences Institute (HESI)
Technical Committee on the Application of Genomics to Mechanism-Based Risk Assessment:

- Steering Committee (5/02-present).
- Hepatotoxicity Working Group Vice-Chair (5/02-present).
- Hepatotoxicity Work Group Member (5/01-present).

Charter member, Fertility and Early Pregnancy Work Group of the National Children's Study (07/01-Present).

National Health and Environmental Effects Research Laboratory Distinguished Lecture Series Committee (July 03-present).

U.S. Environmental Protection Agency Genomics Task Force Microarray Technical Subcommittee (August 03-present).

National Health and Environmental Effects Research Laboratory Genomics and Proteomics Committee (NGPC) (September 03-present).

Professional Meetings:

Invited participant ("Observer") in Expert Panel Workshop: "The Role of Environmental Factors on the Onset and Progression of Puberty in Children". Organised by Serono Symposia International. November 6th-8th, 2003, Chicago, IL, USA.

Joint organiser and co-chair of: "*Genomic analysis of surrogate tissues for measuring toxic exposures and drug action*", the "*Innovations in Applied Toxicology*" Symposium for the Society of Toxicology 42nd Annual Meeting, March 9th-13th, 2003, Salt Lake City, UT, USA.

- (8) **John C. Rockett**, David J. Esdaile and G Gordon Gibson (1999). Differential gene expression in drug metabolism: practicalities, problems and potential. *Xenobiotica*, 29(7):655-691.
- (7) MC Murphy, CN Brookes, **JC Rockett**, C Chapman, JA Lovegrove, BJ Gould, JW Wright and CM Williams (1999). The quantitation of lipoprotein lipase mRNA in biopsies of human adipose tissue, using the polymerase chain reaction, and the effect of increased consumption of n-3 polyunsaturated fatty acids. *European Journal of Clinical Nutrition*, 53:441-447.
- (6) **JC Rockett**, DJ Esdaile and GG Gibson (1997). Molecular profiling of non-genotoxic carcinogenesis using differential display reverse transcription polymerase chain reaction (ddRT-PCR). *European Journal of Drug Metabolism & Pharmacokinetics* 22(4):329-33.
- (5) **Rockett, J.**, Larkin, K., Darnton, S., Morris, A. and Matthews, H. (1997). Five newly established oesophageal carcinoma cell lines: phenotypic and immunological characterisation. *British Journal of Cancer* 75(2):258-263.
- (4) **J C Rockett**, S J Darnton, J Crocker, H R Matthews and A G Morris (1996). Lymphocyte infiltration in oesophageal carcinoma: lack of correlation with MHC antigens, ICAM-1, and tumour stage and grade. *Journal of Clinical Pathology* 49:264-267.
- (3) **J C Rockett**, S J Darnton, J Crocker, H R Matthews and A G Morris (1995). Expression of HL-ABC and HLA-DR histocompatibility antigens and intercellular adhesion molecule-1 in oesophageal carcinoma. *Journal of Clinical Pathology* 48:539-44.
- (2) Salam M, **Rockett J** and Morris A (1995). The prevalence of different human papillomavirus types and p53 mutations in laryngeal carcinomas: is there a reciprocal relationship? *European Journal of Surgical Oncology* 21:290-296.
- (1) Salam M, **Rockett J** and Morris A (1995). General primer-mediated polymerase chain reaction for simultaneous detection and typing of HPV in laryngeal carcinomas. *Clinical Otolaryngology* 20:84-88.

(2) Articles Submitted To A Scientific Journal

- (4) **John C. Rockett**, Judith E. Schmid, Christopher J. Luft, J. Brian Garges, M. Stacey Ricci, Pasquale Patrizio, Norman B. Hecht and David J. Dix. Gene Expression Patterns Associated with Infertility in Rodent and Human Models. *An invited submission*
- (3) Roger Ulrich, **John C. Rockett**, G. Gordon Gibson and Syril Pettit. Evaluating the Effects of Methapyrilene and Clofibrate on Hepatic Gene Expression: A Collaboration Between Laboratories and a Comparison of Platform and Analytical Approaches.
- (2) Valerie A Baker, Helen M Harries, Jeffrey F Waring, Roger Jolly, Angus de Souza, Judith E Schmid, Hong Ni, Roger Brown, Roger G Ulrich and **John C. Rockett**. Clofibrate-Induced Gene Expression Changes in Rat Liver: A Cross-Laboratory Analysis Using Membrane cDNA Arrays.

(1) David Miller, Corrado Spadafora, David Dix, Adrian Platts, **John C. Rockett**, Stephen A Krawetz Nuclease digestion of sperm chromatin suggests a random distribution of gene sequences.

(3) Articles In Preparation For Submission To A Scientific Journal

(3) Spearow J, DB Tully, **John C. Rockett** and DJ Dix. Differential testicular gene expression in mouse strains sensitive and resistant to endocrine disruption by estrogen.

(2) Sally D. Perrault, **John C. Rockett**, Laura Fenster, James Kesner, Wendy Robbins and Steven Schrader. Biomarkers for Assessing Reproductive Development and Health: Part 2 – Adult Reproductive Health.

(1) J. Christopher Luft, Douglas B. Tully, **John C. Rockett**, Judith E. Schmid and David J. Dix. Reproductive and genomic effects in testes from mice exposed to the water disinfectant byproduct bromochloroacetic acid

(4) Book Chapters

(4) **John C. Rockett**. Gene Microarrays Applied to Reproductive Toxicology. In Cunningham (Ed): *Genetic and Proteomic Applications in Toxicity Testing*, The Human Press, Totowa. In Preparation. *An invited submission*

(3) **John C. Rockett** and David J Dix. Gene Expression Networks. In Cooper (ed-in-chief): *Encyclopaedia of the Human Genome*, Nature Publishing Group. London, New York. ISBN 0-333-80386-8 (2003). *An invited submission*

(2) **John C. Rockett**. The Future of Toxicogenomics. In Michael Burczynski (ed): “*An Introduction to Toxicogenomics*”. CRC Press. Boca Raton, London, New York, Washington D.C., pp299-317 (2003). *An invited submission*

(1) **J. Rockett**, S. Darnton, J. Crocker, H. Matthews and A. Morris: Major Histocompatibility Complex (MHC) class I and II and Intercellular Adhesion Molecule (ICAM)-1 expression in oesophageal carcinoma. Peracchia A, Rosati R, Bonavina L, Bona S, Chella B (eds): *Recent Advances in Diseases of the Esophagus*. Bologna: Monduzzi Editore, pp45-49 (1996).

(5) Other Scientific Publications (Letters to Editors; Meeting Reports; Commentaries etc.)

(11) **John C. Rockett** (2003). Probing the nature of microarray-based oligonucleotides. *Drug Discovery Today* 8(9):389. (A Letter To The Editor) *An invited submission*

(10) **John C. Rockett** (2003). To confirm or not to confirm (microarray data) – that, is the question. *Drug Discovery Today* 8(8):343. (A Letter To The Editor)

(9B) Nazzareno Ballatori, James L. Boyer, and **John C. Rockett**. (2003). Exploiting Genome Data to Understand the Function, Regulation and Evolutionary Origins of Toxicologically Relevant Genes. *Environ Health Perspect.* 111(6):871-5. (A Meeting Report)

(9A) Nazzareno Ballatori, James L. Boyer, and **John C. Rockett**. (2003). Exploiting Genome Data to Understand the Function, Regulation and Evolutionary Origins of Toxicologically Relevant Genes. *EHP Toxicogenomics*. 111(1T):61-5. (A Meeting Report)

(8) **John C. Rockett** (2002). Surrogate Tissue Analysis for Monitoring the Degree and Impact of Exposures in Agricultural Workers. *AgBiotechNet*, 4:1-7 November, ABN 100. (A Review Article).
An invited submission

(7) **John C. Rockett** (2002). Macroresults Through Microarrays. *Drug Discovery Today*, 7(15):804-805. (A Meeting Report)

(6) **John C. Rockett** (2002). Chip, chip, array! Three chips for post-genomic research. *Drug Discovery Today*, 7(8):458-459. (A Meeting Report)

(5) **John C. Rockett** (2002). Use of Genomic Data in Risk Assessment. *GenomeBiology*, 3(4):reports4011.1-4011.3 (<http://genomebiology.com/2002/3/4/reports/4011/?isguard=1>). (A Meeting Report)

(4) **John C. Rockett** (2001). Genomic and Proteomic Techniques Applied to Reproductive Biology. *GenomeBiology* 2(9): 4020.1-4020.3 (<http://genomebiology.com/2001/2/9/reports/4020/>). (A Meeting Report)

(3) **John C. Rockett** (2001). Chipping away at the mystery of drug responses. *The Pharmacogenomics Journal*, 1(3):161-163. (A commentary) ***An invited submission***

(2) **Rockett, John C.** and Dix, David J. (1999). U.S. EPA workshop: Application of DNA arrays to Toxicology. *Environmental Health Perspectives*, 107(8):681-685. (A Meeting Report)

(1) **John C. Rockett III** (1995). Immune recognition molecules and transforming growth factor beta-1 in oesophageal cancer. Ph.D. thesis, University of Warwick, Coventry, England. (Ph.D. thesis)

(6) Published Book, Paper and Website reviews

(9) **John C. Rockett** (2002). A report on the manuscript: Systemic RNAi in *C. elegans* requires the putative transmembrane protein SID-1. Winston WM, Molodowitch C, Hunter CP. *Science*. 2002 295:2456-2459. *GenomeBiology*, 3(7):reports0034
<http://genomebiology.com/2002/3/7/reports/0034/>

- (8) **John C. Rockett** (2001). A report on the manuscript: Genetic rescue of an endangered mammal by cross-species nuclear transfer using post-mortem somatic cells. P Loi , et al., *Nat Biotechnol.* 2001, 19:962-964. *GenomeBiology*, 3(1):reports0006. (<http://genomebiology.com/2001/3/1/reports/0006/>).
- (7) **John C. Rockett** (2001). A report on the manuscript: Molecular Classification of Human Carcinomas by Use of Gene Expression Signatures. A Su et al., *Cancer Res.* 2001 61:7388-7393. *GenomeBiology*, 3(1):reports0005. (<http://genomebiology.com/2001/3/1/reports/0005/>).
- (6) **John C. Rockett** (2001). A report on the manuscript: Genetic evidence for two species of elephant in Africa. A Roca et al., *Science*. 2001 Aug 24;293(5534):1473-7. *GenomeBiology*, 2(12):reports0045. (<http://www.genomebiology.com/2001/2/12/reports/0045/>).
- (5) **John C. Rockett** (2001). A report on the manuscript: Extensive genetic polymorphism in the human CYP2B6 gene with impact on expression and function in human liver. T Lang et al., *Pharmacogenetics*, 2001, 11(5):399-415. *GenomeBiology*, 2(12):reports0044. (<http://www.genomebiology.com/2001/2/12/reports/0044/>).
- (4) **John C. Rockett** (2001). A report on the manuscript: Novel Human Testis-Specific cDNA: molecular Cloning, Expression and Immunological Effects of the Recombinant Protein. R Santhanam and R K Naz, *Molecular Reproduction and Development* 60:1-12 (2001). *GenomeBiology*, 2(11):reports0040. (<http://genomebiology.com/2001/2/11/reports/0040/>).
- (3) **John C. Rockett** (2001). A report on the website: BIND - The Biomolecular Interaction Network Database (<http://www.bind.ca/>). *GenomeBiology*, 2(9): reports2011. (<http://www.genomebiology.com/2001/2/9/reports/2011/>).
- (2) **John C. Rockett** (2001). A report on the manuscript: Exploring the DNA-binding specificities of zinc fingers with DNA microarrays. ML Bulyk et al., *Proc Natl Acad Sci USA* 2001, 98:7158-7163. *GenomeBiology*, 2(10): reports0032. (<http://genomebiology.com/2001/2/10/reports/0032/>).
- (1) **J Rockett** (1996). A Book Review on: "Cell Adhesion and Cancer" (Eds., Hogg N. and Hart I.). *Clinical Molecular Pathology* 49(1):M64. *An invited submission*

(7) Published Abstracts of Poster and Oral Presentations

- (17) Amber K. Goetz, Wenjun Bao, Judith E. Schmid, Carmen Wood, Hongzu Ren, Deborah S. Best, Rachel N. Murrell, **John C. Rockett**, Michael G. Narotsky, Douglas C. Wolf, Douglas B. Tully, David J. Dix' Gene Expression Profiling in Testis and Liver of Mice to Identify Modes of Action of Conazole Toxicities. Society of Toxicology 43rd Annual Meeting, March 21st-25th, 2004, Baltimore, MD, USA. *Toxicological Sciences*. (Submitted)
- (16) Jane Gallagher, Theresa Lehman, Ramakrishna Modali, Scott Rhoney, Marien Clas, Jeff Inmon, **John C. Rockett**, David Dix, Cindy Mamay, Suzanne Fenton, Suzanne McMaster, Stan

Barone Jr, Pauline Mendola and Reeder Sams. Validation of Non-Invasive Biological Samples: Pilot Projects Relevant to the National Children Study. Society of Toxicology 43rd Annual Meeting, March 21st-25th, 2004, Baltimore, MD, USA. *Toxicological Sciences*. (Submitted)

(15) B.S. Pukazhenth, **J. C. Rockett**, M. Ouyang, D.J. Dix, J.G. Howard, P. Georgopoulos, W.J. J. Welsh and D. E. Wildt. Gene Expression In The Testis Of Normospermic Versus Teratospermic Domestic Cats Using Human cDNA Microarray Analyses. Society for the Study of Reproduction 36th Annual Meeting, July 19th-22nd, 2003, Cincinnati, OH, USA. *Biology of Reproduction* 68 (Supp 1):191.

(14) David J. Dix and **John C. Rockett** (2003). Genomic and Proteomic Analysis of Surrogate Tissues for Assessing Toxic Exposures and Disease States. Innovation in Applied Toxicology symposium entitled "*Genomic and Proteomic Analysis of Surrogate Tissues for Assessing Toxic Exposures and Disease States*". Society of Toxicology 42nd Annual Meeting, March 9th-13th, 2003, Salt Lake City, UT, USA. *Toxicological Sciences* 72(S-1):276.

(13) **John C. Rockett**, Chad R. Blystone, Amber K. Goetz, Rachel N. Murrell, Judith E. Schmid and David J. Dix. (2003). Gene Expression Profiling Of Accessible Surrogate Tissues To Monitor Molecular Changes In Inaccessible Target Tissues Following Toxicant Exposure. Innovations in Applied Toxicology Symposium entitled "*Genomic and Proteomic Analysis of Surrogate Tissues for Assessing Toxic Exposures and Disease States*". Society of Toxicology 42nd Annual Meeting, March 9th-13th, 2003, Salt Lake City, UT, USA. *Toxicological Sciences* 72(S-1):276.

(12) Douglas B. Tully, J. Christopher Luft, **John C. Rockett**, Judy E. Schmid and David J. Dix (2002). Effects on gene expression in testes from adult male mice exposed to the water disinfectant byproduct bromochloroacetic acid. *Society for the Study of Reproduction 35th Annual Meeting, July 28-31, 2002, Baltimore, Maryland, USA. Biology of Reproduction* 66 (Supp 1):223.

(11) David J. Dix, Kary E. Thompson, **John C. Rockett**, Judith E. Schmid, Robert J. Goodrich, David Miller, G. Charles Ostermeier and Stephen A. Krawetz (2002). Testis and spermatzoa RNA profiles of normal fertile men. *Society for the Study of Reproduction 35th Annual Meeting, July 28-31, 2002, Baltimore, Maryland, USA. Biology of Reproduction* 66 (Supp 1):194.

(10) Asa J. Oudes, **John C. Rockett**, David J. Dix and Kwan Hee Kim (2002). Identification of retinoic acid induced genes in mouse testis by cDNA microarray analysis. *27th Annual Meeting of the American Society of Andrology, 4/24-27/02. J. Andrology Supplement March/April*.

(9) **John C. Rockett**, Robert J. Kavlock, Christy Lambright, Louise G. Parks, Judith E. Schmid, Vickie S. Wilson and David J. Dix (2002). Use of DNA arrays to monitor gene expression in blood and uterus from Long-Evans rats following 17- β -estradiol exposure – a new approach to biomonitoring endocrine disrupting chemicals using surrogate tissues. *Toxicological Sciences* 66(1): Abstract No.1388.

(8) David J. Dix and **John C. Rockett** (2002). Genomic analysis of the testicular toxicity of haloacetic acids. Platform presentation at the symposium, "Defining the cellular and molecular

mechanisms of toxicant action in the testis". *Toxicological Science* 66 (1): Abstract No.848.

(7) **JC Rockett**, JC Luft, JB Garges and DJ Dix (2001). The reproductive effects of the water disinfectant byproduct bromochloroacetate on juvenile and adult male mice. *Toxicological Sciences*, 60 (1):250.

(6) Tarka DK, Klinefelter GR, **Rockett JC**, Suarez JD, Roberts NL and Rogers JM (2001). Effect of gestational exposure to ethane dimethane sulfonate (EDS), bromochloroacetic acid (BCA) and molinate on reproductive function in CD-1 male mice. *Toxicological Sciences*, 60 (1):250.

(5) Garges JB, **Rockett JC** and Dix DJ (2001). Developmental and reproductive phenotype of mice lacking stress-inducible 70 kDa heat shock proteins (Hsp70s). *Toxicological Sciences*, 60 (1):383.

(4) D Dix, **J Rockett**, J Luft, J Garges, M Ricci, P Patrizio and N Hecht (2000). Using DNA microarrays to characterise gene expression in testes of fertile and infertile humans and mice. *Biology of Reproduction*, 62 (s1):227.

(3) J Luft, J B Garges, **J Rockett** and D Dix (2000). Male reproductive toxicity of bromochloroacetic acid in mice. *Biology of Reproduction*, 62 (s1):246.

(2) **Rockett, JC**, Garges, JB and Dix, DJ (2000). A single heat-shock of juvenile male mice causes a long-term decrease in fertility and reduces embryo quality. *Toxicological Sciences* 54 (1):365.

(1) **JC Rockett**, SJ Darnton, J Crocker, HR Matthews and AG Morris (1994). Major Histocompatibility (MHC) class I and II and intercellular adhesion molecule (ICAM)-1 expression in oesophageal carcinoma (OC). *Immunology* 83 (s1):64.

(8) Invited Oral Presentations

(10) **John C. Rockett** and Gary M Hellmann. *To confirm or not to confirm (microarray data) – that is the question*. Seminar for EPA/NHEERL Genomics and Proteomics Committee's ArrayQA forum, August 25th, 2003, RTP, NC, USA.

(9) **John C. Rockett**. *"Biomonitoring Toxicant Exposure and Effect Using Toxicogenomics and Surrogate Tissue Analysis"*. Seminar for Division of Epidemiology, Statistics and Prevention Research, National Institute of Child Health and Development, May 29th, 2003, Rockville, MD, USA.

(8) **John C. Rockett**. *"Genomics and Proteomics: New Toxicity Testing"*. Platform presentation at US EPA Regional Risk Assessors Annual Conference, April 28th – May 2nd, 2003, Stone Mountain, GA, USA.

(7) **John C. Rockett**, Chad R. Blystone, Amber K. Goetz, Rachel N. Murrell, Judith E. Schmid and David J. Dix. *"Gene Expression Profiling Of Accessible Surrogate Tissues To Monitor Molecular Changes in Inaccessible Target Tissues Following Toxicant Exposure."* Platform presentation at

SoT 42nd Annual Meeting symposium entitled "*Genomic and Proteomic Analysis of Surrogate Tissues for Measuring Toxic Exposures and Drug Action*", March 9th-13th, 2003, Salt Lake City, UT, USA.

(6) **John C. Rockett**. "*A Toxicogenomic Approach to Surrogate Tissue Analysis*". Seminar for Department of Environmental and Molecular Toxicology, North Carolina State University, September 3rd, 2002, Raleigh, NC, USA.

(5) **John C. Rockett**. "Differential gene expression in toxicology: practicalities, problems and potential". Platform presentation at *9th Annual Mount Desert Island Biological Laboratory Environmental Health Sciences Symposium: Exploiting Genome Data to Understand the Function, Regulation and Evolutionary Origins of Toxicologically Relevant Genes*, July 10th-11th, 2002, Salisbury Cove, Maine, USA.

(4) **John C. Rockett**, Leroy Folmar, Michael J. Hemmer and David J. Dix. "Arrays for biomonitoring environmental and reproductive toxicology". Platform Presentation at *Macroresults Through Microarrays 3 – Advancing Drug Development*, April 29th-May 1st, 2002, Boston, MA, USA.

(3) **John C. Rockett**, Sigmund Degitz, Suzanne E. Fenton, Leroy Folmar, Michael J. Hemmer, Joe E Tietge, and David J. Dix. "Use of DNA Arrays in Environmental Toxicology". Platform presentation at the *4th Annual Lab-on-a-Chip and Microarrays for Post-Genomic Applications meeting*, January 14th-16th, 2002, Zurich, Switzerland.

(2) **John C. Rockett**. "DNA Arrays". Seminar at *EPA Molecular Biology Course*, April 8th, 1999, USEPA, RTP, NC, USA.

(1) **John C. Rockett**. "Contract Services for Array Applications". Seminar at the *Triangle Array Users Group*, May 1st, 1999, CIIT, RTP, NC, USA.

(9) Other Poster and Oral Presentations

(23) **John C. Rockett**, Wenjun Bao, Chad R. Blystone, Amber K. Goetz, Rachel N. Murrell, Hongzu Ren, Judith E. Schmid, Jessica Stapelfeldt, Lillian F. Strader, Kary E. Thompson and David J. Dix. Genomic Analysis of Surrogate Tissues for Assessing Environmental Exposures and Future Disease States. ILSI-HESI meeting: *Toxicogenomics in Risk Assessment - Assessing the Utility, Challenges, and Next Steps*. June 5th-6th, 2003, Fairfax, VA, USA.

(22) **John C. Rockett**, Wenjun Bao, Chad R. Blystone, Amber K. Goetz, Rachel N. Murrell, Hongzu Ren, Judith E. Schmid, Jessica Stapelfeldt, Lillian F. Strader, Kary E. Thompson and David J. Dix. Genomic Analysis of Surrogate Tissues for Assessing Environmental Exposures and Future Disease States. *EPA Science Forum*, May 5th-7th, 2003, Washington, D.C., USA.

(21) Germaine Buck, Courtney Johnson, Joseph Stanford, Anne Sweeney, Laura Schieve, **John Rockett**, Sherry Selevan and Steve Schrader. Prospective Pregnancy Study Designs for Assessing Reproductive and Developmental Toxicants. *American Epidemiology Society Meeting*, March 27th-28th, 2003, Atlanta, GA, USA.

(20) **John C. Rockett**, Chad R. Blystone, Amber K. Goetz, Rachel N. Murrell, Hongzu Ren, Judith E. Schmid, Jessica Stapelfeldt, Lillian F. Strader, Kary E. Thompson, Doug B. Tully, Paul Zigas and David J. Dix. Genomic Analysis of Surrogate Tissues for Assessing Environmental Exposures and Future Disease States. *National Children's Study Assembly Meeting*, December 16th-18th, 2002, Baltimore, MD, USA.

(19) **John Rockett**. The Use of Gene Expression Profiling to Detect Early Biomarkers of Adverse Effects Prior to Clinical manifestation. *National Children's Study: Meeting of EPA Project Leaders – Methods Development Projects*. November 20th, 2002, USEPA, RTP, NC, USA. (Oral Presentation)

(18) GC Ostermeier, RJ Goodrich, K Thompson, **J Rockett**, MP Diamond, K Collins, NICHD Reproductive Medicine Network, DJ. Dix, D Miller and SA Krawetz. Defining the spermatozoal RNA population in normal fertile men. *American Society of Reproductive Medicine* October 12-17, 2002, Seattle, WA, USA.

(17) G. Charles Ostermeier, Robert J. Goodrich, Kary Thompson, **John Rockett**, Michael P. Diamond, Karen Collins, NICHD Reproductive Medicine Network, David J. Dix, David Miller and Stephen A. Krawetz. RNAs isolated from ejaculate spermatozoa provide a noninvasive means to investigate testicular gene expression. *Gordon Conference on Mammalian Gametogenesis & Embryogenesis*, June 30th-July 5th, Connecticut College, New London, CT, USA.

(16) David Dix, **John Rockett**, Judith Schmid, Lillian Strader, Douglas Tully. Genomic analysis of testicular toxicity. *USEPA/NHEERL/RTD Peer Review*, October 22nd, 2001, RTP, NC, USA.

(15) David Dix, **John Rockett**, Judith Schmid, Douglas Tully. Monitoring human reproductive health and development through gene expression profiling. *USEPA/NHEERL/RTD Peer Review*, October 22nd, 2001, RTP, NC, USA.

(14) Patrizio P, N Hecht, **J Rockett**, J Schmid and D Dix (2001). DNA microarrays to study gene expression profiles in testis of fertile and infertile men. *57th Annual Meeting of the American Society for Reproductive Medicine*, October 20th-25th, 2001, Orlando, FL, USA.

(13) Jimmy L. Spearow, Dale Morris, Uland Wong, Rashid Altafi, Saeed Eteawi, Mark Stanford, Trevor Stearns, Lorena Orozco, Angela Chen, **John Rockett**, Douglas Tully, David Dix and Marylynn Barkley. Genetic Variation In Susceptibility To The Disruption Of Testicular Development And Gene Expression By Pubertal Exposure To Estrogenic Agents. *Third Annual University of California at Davis Conference for Environmental Health Scientists, Disruption of Developing Systems and Advances in Therapeutic Approaches* August 27th, 2001, UC Davis, CA, USA.

(12) Tarka DK, Klinefelter GR, **Rockett JC**, Suarez JD, Roberts NL and Rogers JM (2001). Effect of gestational exposure to ethane dimethane sulfonate (EDS), bromochloroacetic acid (BCA) and molinate on reproductive function in CD-1 male mice. *North Carolina Society of Toxicology Winter Meeting*, March 3rd, 2001. NIEHS, RTP, NC, USA.

(11) David Dix, **John Rockett**, Leroy Folmar, Michael Hemmer, Sigmund Degitz, and Joseph Tietge (2001). Biomonitoring the Toxicogenomic Response to Endocrine Disrupting Chemicals in Humans, Laboratory Species and Wildlife. *U.S. - Japan International Workshop for Endocrine Disrupting Chemicals*, February 28th-March 3rd, 2001, Tsukuba, Japan.

(10) **John C. Rockett**, Faye L. Mapp, J. Brian Garges, J. Christopher Luft, Chisato Mori and David J Dix (2001). The effects of hyperthermia on spermatogenesis, apoptosis, gene expression and fertility in adult male mice. *Triangle Consortium for Reproductive Biology Annual Meeting*, January 27th, 2001, RTP, NC, USA.

(9) Gangolli E, Dix DJ, Garges J B, **Rockett, JC** and Idzerda RL (2000). Testosterone Regulation of Sertoli Cell genes. *11th International Congress of Endocrinology*, October 29th-November 2nd, 2000, Sydney, Australia.

(8) **J Rockett**, J Luft, J Garges, M Ricci, P Patrizio, N Hecht and D Dix (2000). Using DNA microarrays to characterise gene expression in testes of fertile and infertile humans and mice. *Functional Genomics & Microarray Data Mining*, August 3rd-4th 2000, Durham, NC, USA.

(7) **Rockett JC**, S Ricci, P Patrizio, NB Hecht, JB Garges and DJ Dix (2000). Gene Expression in the Mammalian Testis. *5th NHEERL Symposium*, June 6th-8th, 2000, RTP, NC, USA.

(6) J Luft, J B Garges, **J Rockett** and D Dix (2000). Male reproductive toxicity of bromochloroacetic acid in mice. *2000 NIEHS/NTA Biomedical Science and Career Fair*, April 28th 2000, RTP, NC, USA.

(5) **Rockett JC**, S Ricci, P Patrizio, NB Hecht, JB Garges and DJ Dix (2000). Gene Expression in the Mammalian Testis. *Molecular Toxicology, Toxicogenomics and Associated Bioinformatics Applied to Drug Discovery meeting*, January 11th-15th, 2000, Santa Fe, NM, USA.

(4) **JC Rockett** and DJ Dix (1999). Development of DNA arrays for the analysis of testis-expressed genes in humans and mice. *The 8th Annual National Health and Environmental Effects Research Laboratory Open House*. November 2nd-3rd, 1999, RTP, NC, USA.

(3) **JC Rockett**, DJ Esdaile and GG Gibson (1997). Molecular profiling of non-genotoxic carcinogenesis using differential display reverse transcription polymerase chain reaction (ddRT-PCR). *The British Toxicology Society Annual Meeting*, April 19th-22nd, 1998, University of Surrey, Guildford, Surrey, England.

(2) **JC Rockett**, DJ Esdaile and GG Gibson (1997). Molecular profiling of non-genotoxic

carcinogenesis using differential display reverse transcription polymerase chain reaction (ddRT-PCR). Poster presentation at *Symposium on Drug Metabolism: Towards the next Millennium*. August 26th-28th, 1997, London King's College, London, England.

(1) **J Rockett**, S Darnton, J Crocker, H Matthews and A Morris: Major Histocompatibility Complex (MHC) class I and II and Intercellular Adhesion Molecule (ICAM)-1 expression in oesophageal carcinoma. Oral presentation at *The 6th World Congress of the International Society for Diseases of the Esophagus*, August 23rd-26th, 1995, Milan, Italy.

uct. pC225 is a KS+ (Stratagene) plasmid containing a 0.5-kb Bam HI-Sst I fragment from pAXL1. Substitution mutations of the proposed active site of Axl1p were created with the use of pC225 and site-specific mutagenesis involving appropriate synthetic oligonucleotides (axl1-H68A, 5'-GTGCTCACAAGGCGT-GCCAAACCGGC-3'; axl1-E71A, 5'-AAGAATCATGTGCGCACAAAGGTGCGC-3'; and axl1-E71D, 5'-AAGAATCATGTGATCACAAGGTGCGC-3'). The mutations were confirmed by sequence analysis. After mutagenesis, the 0.4-kb Bam HI-Msc I fragment from the mutagenized pC225 plasmids was transferred into pAXL1 to create a set of pRS316 plasmids carrying different AXL1 alleles, p124 (axl1-H68A), p130 (axl1-E71A), and p132 (axl1-E71D). Similarly, a set of HA-tagged alleles carried on YEp352 were created after replacement of the p151 Bam HI-Msc I fragment, to generate p161 (axl1-E71A), p162 (axl1-

N. Davis, T. Favero, C. de Hoog, and S. Kim for comments on the manuscript. Supported by a grant to C.B. from the Natural Sciences and Engineering Research Council of Canada. Support for M.N.A. was from a California Tobacco-Related Disease Research Program postdoctoral fellowship (4FT-0083).

22 June 1995; accepted 21 August 1995

Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray

Mark Schena,* Dari Shalon,*† Ronald W. Davis,
Patrick O. Brown‡

A high-capacity system was developed to monitor the expression of many genes in parallel. Microarrays prepared by high-speed robotic printing of complementary DNAs on glass were used for quantitative expression measurements of the corresponding genes. Because of the small format and high density of the arrays, hybridization volumes of 2 microliters could be used that enabled detection of rare transcripts in probe mixtures derived from 2 micrograms of total cellular messenger RNA. Differential expression measurements of 45 *Arabidopsis* genes were made by means of simultaneous, two-color fluorescence hybridization.

The temporal, developmental, topographical, histological, and physiological patterns in which a gene is expressed provide clues to its biological role. The large and expanding database of complementary DNA (cDNA) sequences from many organisms (1) presents the opportunity of defining these patterns at the level of the whole genome.

For these studies, we used the small flowering plant *Arabidopsis thaliana* as a model organism. *Arabidopsis* possesses many advantages for gene expression analysis, including the fact that it has the smallest genome of any higher eukaryote examined to date (2). Forty-five cloned *Arabidopsis* cDNAs (Table 1), including 14 complete sequences and 31 expressed sequence tags (ESTs), were used as gene-specific targets. We obtained the ESTs by selecting cDNA clones at random from an *Arabidopsis* cDNA library. Sequence analysis revealed that 28 of the 31 ESTs matched sequences

in the database (Table 1). Three additional cDNAs from other organisms served as controls in the experiments.

The 48 cDNAs, averaging ~1.0 kb, were amplified with the polymerase chain reaction (PCR) and deposited into individual wells of a 96-well microtiter plate. Each sample was duplicated in two adjacent wells to allow the reproducibility of the arraying and hybridization process to be tested. Samples from the microtiter plate were printed onto glass microscope slides in an area measuring 3.5 mm by 5.5 mm with the use of a high-speed arraying machine (3). The arrays were processed by chemical and heat treatment to attach the DNA sequences to the glass surface and denature them (3). Three arrays, printed in a single lot, were used for the experiments here. A single microtiter plate of PCR products provides sufficient material to print at least 500 arrays.

Fluorescent probes were prepared from total *Arabidopsis* mRNA (4) by a single round of reverse transcription (5). The *Arabidopsis* mRNA was supplemented with human acetylcholine receptor (AChR) mRNA at a dilution of 1:10,000 (w/w) before cDNA synthesis, to provide an internal standard for calibration (5). The resulting fluorescently labeled cDNA mixture was hybridized to an array at high stringency (6) and scanned

M. Schena and R. W. Davis, Department of Biochemistry, Beckman Center, Stanford University Medical Center, Stanford, CA 94305, USA.
D. Shalon and P. O. Brown, Department of Biochemistry and Howard Hughes Medical Institute, Beckman Center, Stanford University Medical Center, Stanford, CA 94305, USA.

*These authors contributed equally to this work.

†Present address: Syntent, Palo Alto, CA 94303, USA.

‡To whom correspondence should be addressed. E-mail: pbrown@cmgm.stanford.edu

- Axl1p sequence following Ser²⁰⁶ and occurs within the domain of Axl1p that shows homology with HDE (14). To delete the complete STE23 sequence and create the ste23Δ::URA3 mutation, polymerase chain reaction (PCR) primers (5'-TCGGAGACCTCAT-TCTTGCTCATTTGATATTGCTC- TGATAGATTG-TACTGAGAGTGCAC-3'; and 5'-GCTACAACAGC-GTGGACTTGAATGCCCGGACATCTTGGACTGT-GGGGTATTTCACACG-3') were used to amplify the URA3 sequence of pRS316, and the reaction product was transformed into yeast for one-step gene replacement [R. Rothstein, *Methods Enzymol.* 194, 281 (1991)]. To create the axl1Δ::LEU2 mutation contained on p114, a 5.0-kb Sal I fragment from pAXL1 was cloned into pUC19, and an internal 4.0-kb Hpa I-Xho I fragment was replaced with a LEU2 fragment. To construct the ste23Δ::LEU2 allele (a deletion corresponding to 931 amino acids) carried on p153, a LEU2 fragment was used to replace the 2.8-kb Pml I-Ecl136 II fragment of STE23, which occurs within a 6.2-kb Hind III-Bgl II genomic fragment carried on pSP72 (Promega). To create YEpMFA1, a 1.8-kb Bam HI fragment containing MFA1, from pKK16 [K. Kuchler, R. E. Sterne, J. Thorer, *EMBO J.* 8, 3973 (1989)], was ligated into the Bam HI site of YEp351 [J. E. Hill, A. M. Myers, T. J. Koerner, A. Tzagioff, *Yeast* 2, 163 (1986)].
24. J. Chant and I. Herskowitz, *Cell* 65, 1203 (1991).
 25. B. W. Matthews, *Acc. Chem. Res.* 21, 333 (1988).
 26. K. Kuchler, H. G. Dohlman, J. Thorer, *J. Cell Biol.* 120, 1203 (1993); R. Kolling and C. P. Hollenberg, *EMBO J.* 13, 3261 (1994); C. Berkower, D. Loeyza, S. Michaelis, *Mol. Biol. Cell* 5, 1185 (1994).
 27. A. Bender and J. R. Pringle, *Proc. Natl. Acad. Sci. U.S.A.* 86, 9976 (1989); J. Chant, K. Corrado, J. R. Pringle, I. Herskowitz, *Cell* 65, 1213 (1991); S. Powers, E. Gonzales, T. Christensen, J. Cubert, D. Broek, *ibid.*, p. 1225; H. O. Park, J. Chant, I. Herskowitz, *Nature* 365, 269 (1993); J. Chant, *Trends Genet.* 10, 328 (1994); — and J. R. Pringle, *J. Cell Biol.* 129, 751 (1995); J. Chant, M. Mischke, E. Mitchell, I. Herskowitz, J. R. Pringle, *ibid.*, p. 767.
 28. G. F. Sprague Jr., *Methods. Enzymol.* 194, 77 (1991).
 29. Single-letter abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.
 30. A W303 1A derivative, SY2625 (MATa ura3-1 leu2-3, 112 trp1-1 ade2-1 can1-100 ss11Δ ma2Δ::FUS1-lacZ his3Δ::FUS1-HIS3), was the parent strain for the mutant search. SY2625 derivatives for the mating assays, secreted pheromone assays, and the pulse-chase experiments included the following strains: Y49 (ste22-1), Y115 (ma1Δ::LEU2), Y142 (axl1::URA3), Y173 (axl1Δ::LEU2), Y220 (axl1::URA3 ste23Δ::URA3), Y221 (ste23Δ::URA3), Y231 (axl1Δ::LEU2 ste23Δ::LEU2), and Y233 (ste23Δ::LEU2). MATa derivatives of SY2625 included the following strains: Y199 (SY2625 made MATa), Y278 (ste22-1), Y195 (ma1Δ::LEU2), Y196 (axl1Δ::LEU2), and Y197 (axl1::URA3). The EG123 (MATa leu2 ura3 trp1 can1 his4) genetic background was used to create a set of strains for analysis of bud site selection. EG123 derivatives included the following strains: Y175 (axl1Δ::LEU2), Y223 (axl1::URA3), Y234 (ste23Δ::LEU2), and Y272 (axl1Δ::LEU2 ste23Δ::LEU2). MATa derivatives of EG123 included the following strains: Y214 (EG123 made MATa) and Y293 (axl1Δ::LEU2). All strains were generated by means of standard genetic or molecular methods involving the appropriate constructs (23). In particular, the axl1 ste23 double mutant strains were created by crossing of the appropriate MATa ste23 and MATa axl1 mutants, followed by sporulation of the resultant diploid and isolation of the double mutant from nonparental di-type tetrads. Gene disruptions were confirmed with either PCR or Southern (DNA) analysis.
 31. p129 is a YEp352 [J. E. Hill, A. M. Myers, T. J. Koerner, A. Tzagioff, *Yeast* 2, 163 (1986)] plasmid containing a 5.5-kb Sal I fragment of pAXL1. p151 was derived from p129 by insertion of a linker at the Bgl II site within AXL1, which led to an in-frame insertion of the hemagglutinin (HA) epitope (DQTPYDPPYA) (29) between amino acids 854 and 855 of the AXL1 prod-

with a laser (3). A high-sensitivity scan gave signals that saturated the detector at nearly all of the *Arabidopsis* target sites (Fig. 1A). Calibration relative to the AChR mRNA standard (Fig. 1A) established a sensitivity limit of $\sim 1:50,000$. No detectable hybridization was observed to either the rat glucocorticoid receptor (Fig. 1A) or the yeast TRP4 (Fig. 1A) targets even at the highest scanning sensitivity. A moderate-sensitivity scan

of the same array allowed linear detection of the more abundant transcripts (Fig. 1B). Quantitation of both scans revealed a range of expression levels spanning three orders of magnitude for the 45 genes tested (Table 2). RNA blots (7) for several genes (Fig. 2) corroborated the expression levels measured with the microarray to within a factor of 5 (Table 2).

Differential gene expression was investi-

gated with a simultaneous, two-color hybridization scheme, which served to minimize experimental variation inherent in the comparison of independent hybridizations. Fluorescent probes were prepared from two mRNA sources with the use of reverse transcriptase in the presence of fluorescein- and lissamine-labeled nucleotide analogs, respectively (5). The two probes were then mixed together in equal proportions, hybridized to a single array, and scanned separately for fluorescein and lissamine emission after independent excitation of the two fluorophores (3).

To test whether overexpression of a single gene could be detected in a pool of total *Arabidopsis* mRNA, we used a microarray to analyze a transgenic line overexpressing the single transcription factor HAT4 (8). Fluorescent probes representing mRNA from wild-type and HAT4-transgenic plants were labeled with fluorescein and lissamine, respectively; the two probes were then mixed and hybridized to a single array. An intense hybridization signal was observed at the position of the HAT4 cDNA in the lissamine-specific scan (Fig. 1D), but not in the fluorescein-specific scan of the same array (Fig. 1C). Calibration with AChR mRNA added to the fluorescein- and lissamine cDNA synthesis reactions at dilutions of 1:10,000 (Fig. 1C) and 1:100 (Fig. 1D), respectively, revealed a 50-fold elevation of HAT4 mRNA in the transgenic line relative to its abundance in wild-type plants (Table 2). This magnitude of HAT4 overexpression matched that inferred from the Northern (RNA) analysis within a factor of 2 (Fig. 2 and Table 2). Expression of all the other genes monitored on the array differed by less than a factor of 5 between HAT4-transgenic and wild-type plants (Fig. 1, C

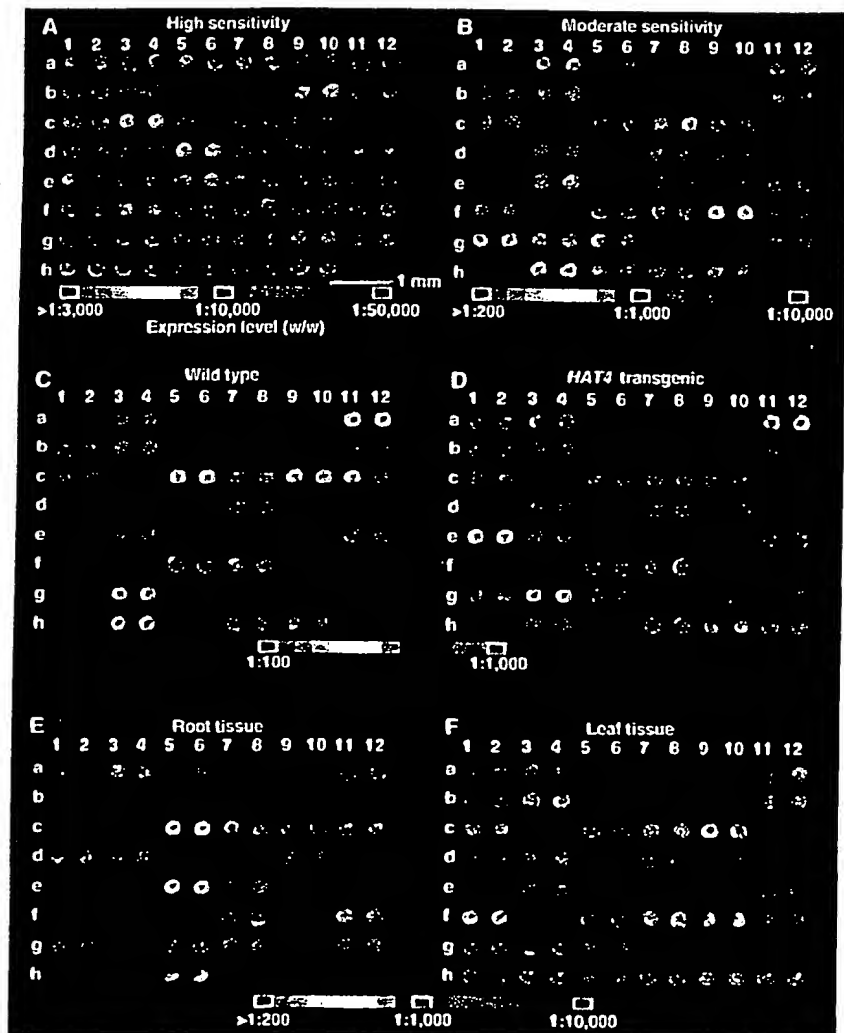


Fig. 1. Gene expression monitored with the use of cDNA microarrays. Fluorescent scans represented in pseudocolor correspond to hybridization intensities. Color bars were calibrated from the signal obtained with the use of known concentrations of human AChR mRNA in independent experiments. Numbers and letters on the axes mark the position of each cDNA. (A) High-sensitivity fluorescein scan after hybridization with fluorescein-labeled cDNA derived from wild-type plants. (B) Same array as in (A) but scanned at moderate sensitivity. (C and D) A single array was probed with a 1:1 mixture of fluorescein-labeled cDNA from wild-type plants and lissamine-labeled cDNA from HAT4-transgenic plants. The single array was then scanned successively to detect the fluorescein fluorescence corresponding to mRNA from wild-type plants (C) and the lissamine fluorescence corresponding to mRNA from HAT4-transgenic plants (D). (E and F) A single array was probed with a 1:1 mixture of fluorescein-labeled cDNA from root tissue and lissamine-labeled cDNA from leaf tissue. The single array was then scanned successively to detect the fluorescein fluorescence corresponding to mRNAs expressed in roots (E) and the lissamine fluorescence corresponding to mRNAs expressed in leaves (F).

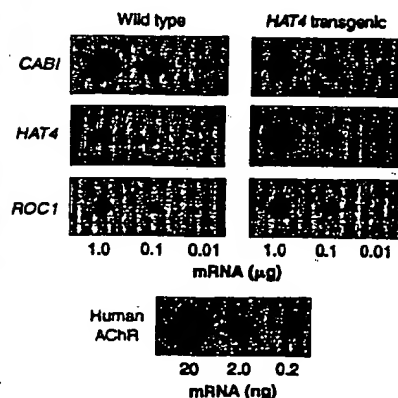


Fig. 2. Gene expression monitored with RNA (Northern) blot analysis. Designated amounts of mRNA from wild-type and HAT4-transgenic plants were spotted onto nylon membranes and probed with the cDNAs indicated. Purified human AChR mRNA was used for calibration.

and D, and Table 2). Hybridization of fluorescein-labeled glucocorticoid receptor cDNA (Fig. 1C) and lissamine-labeled TRP4 cDNA (Fig. 1D) verified the presence of the negative control targets and the lack of optical cross talk between the two fluorophores.

To explore a more complex alteration in expression patterns, we performed a second two-color hybridization experiment with fluorescein- and lissamine-labeled probes prepared from root and leaf mRNA, respectively. The scanning sensitivities for the two fluorophores were normalized by matching the signals resulting from AChR

mRNA, which was added to both cDNA synthesis reactions at a dilution of 1:1000 (Fig. 1, E and F). A comparison of the scans revealed widespread differences in gene expression between root and leaf tissue (Fig. 1, E and F). The mRNA from the light-regulated *CAB1* gene was ~500-fold more abundant in leaf (Fig. 1F) than in root tissue (Fig. 1E). The expression of 26 other genes differed between root and leaf tissue by more than a factor of 5 (Fig. 1, E and F).

The *HAT4*-transgenic line we examined has elongated hypocotyls, early flowering, poor germination, and altered pigmentation (8). Although changes in expression were

observed for *HAT4*, large changes in expression were not observed for any of the other 44 genes we examined. This was somewhat surprising, particularly because comparative analysis of leaf and root tissue identified 27 differentially expressed genes. Analysis of an expanded set of genes may be required to identify genes whose expression changes upon *HAT4* overexpression; alternatively, a comparison of mRNA populations from specific tissues of wild-type and *HAT4*-transgenic plants may allow identification of downstream genes.

At the current density of robotic printing, it is feasible to scale up the fabrication process to produce arrays containing 20,000 cDNA targets. At this density, a single array would be sufficient to provide gene-specific targets encompassing nearly the entire repertoire of expressed genes in the *Arabidopsis* genome (2). The availability of 20,274 ESTs from *Arabidopsis* (1, 9) would provide a rich source of templates for such studies.

The estimated 100,000 genes in the human genome (10) exceeds the number of *Arabidopsis* genes by a factor of 5 (2). This modest increase in complexity suggests that similar cDNA microarrays, prepared from the rapidly growing repertoire of human ESTs (1), could be used to determine the expression patterns of tens of thousands of human genes in diverse cell types. Coupling an amplification strategy to the reverse transcription reaction (11) could make it feasible to monitor expression even in minute tissue samples. A wide variety of acute and chronic physiological and pathological conditions might lead to characteristic changes in the patterns of gene expression in peripheral blood cells or other easily sampled tissues. In concert with cDNA microarrays for monitoring complex expression patterns, these tissues might therefore serve as sensitive *in vivo* sensors for clinical diagnosis. Microarrays of cDNAs could thus provide a useful link between human gene sequences and clinical medicine.

Table 2. Gene expression monitoring by microarray and RNA blot analyses; tg, *HAT4*-transgenic. See Table 1 for additional gene information. Expression levels (w/w) were calibrated with the use of known amounts of human AChR mRNA. Values for the microarray were determined from microarray scans (Fig. 1); values for the RNA blot were determined from RNA blots (Fig. 2).

Gene	Expression level (w/w)	
	Microarray	RNA blot
<i>CAB1</i>	1:48	1:83
<i>CAB1</i> (tg)	1:120	1:150
<i>HAT4</i>	1:8300	1:6300
<i>HAT4</i> (tg)	1:150	1:210
<i>ROC1</i>	1:1200	1:1800
<i>ROC1</i> (tg)	1:260	1:1300

Table 1. Sequences contained on the cDNA microarray. Shown is the position, the known or putative function, and the accession number of each cDNA in the microarray (Fig. 1). All but three of the ESTs used in this study matched a sequence in the database. NADH, reduced form of nicotinamide adenine dinucleotide; ATPase, adenosine triphosphatase; GTP, guanosine triphosphate.

Position	cDNA	Function	Accession number
a1, 2	AChR	Human AChR	-
a3, 4	EST3	Actin	H36236
a5, 6	EST6	NADH dehydrogenase	Z27010
a7, 8	AAC1	Actin 1	M20016
a9, 10	EST12	Unknown	U36594†
a11, 12	EST13	Actin	T45783
b1, 2	<i>CAB1</i>	Chlorophyll a/b binding	M85150
b3, 4	EST17	Phosphoglycerate kinase	T44490
b5, 6	GA4	Gibberellic acid biosynthesis	L37126
b7, 8	EST19	Unknown	U36596†
b9, 10	<i>GBF-1</i>	G-box binding factor 1	X63894
b11, 12	EST23	Elongation factor	X52256
c1, 2	EST29	Aldolase	T04477
c3, 4	<i>GBF-2</i>	G-box binding factor 2	X63895
c5, 6	EST34	Chloroplast protease	R87034
c7, 8	EST35	Unknown	T14152
c9, 10	EST41	Catalase	T22720
c11, 12	rGR	Rat glucocorticoid receptor	M14053
d1, 2	EST42	Unknown	U36596†
d3, 4	EST45	ATPase	J04185
d5, 6	<i>HAT1</i>	Homeobox-leucine zipper 1	U09332
d7, 8	EST46	Light harvesting complex	T04063
d9, 10	EST49	Unknown	T76267
d11, 12	<i>HAT2</i>	Homeobox-leucine zipper 2	U09335
e1, 2	<i>HAT4</i>	Homeobox-leucine zipper 4	M90394
e3, 4	EST50	Phosphoribulokinase	T04344
e5, 6	<i>HAT5</i>	Homeobox-leucine zipper 5	M90416
e7, 8	EST51	Unknown	Z33675
e9, 10	<i>HAT22</i>	Homeobox-leucine zipper 22	U09336
e11, 12	EST52	Oxygen evolving	T21749
f1, 2	EST59	Unknown	Z34607
f3, 4	<i>KNAT1</i>	Knotted-like homeobox 1	U14174
f5, 6	EST60	RuBisCO small subunit	X14564
f7, 8	EST69	Translation elongation factor	T42799
f9, 10	<i>PPH1</i>	Protein phosphatase 1	U34803
f11, 12	EST70	Unknown	T44621
g1, 2	EST75	Chloroplast protease	T43698
g3, 4	EST78	Unknown	R65481
g5, 6	<i>ROC1</i>	Cyclophilin	L14844
g7, 8	EST82	GTP binding	X59152
g9, 10	EST83	Unknown	Z33795
g11, 12	EST84	Unknown	T45278
h1, 2	EST91	Unknown	T13832
h3, 4	EST96	Unknown	R64816
h5, 6	<i>SAR1</i>	Synaptobrevin	M90418
h7, 8	EST100	Light harvesting complex	Z18205
h9, 10	EST103	Light harvesting complex	X03909
h11, 12	<i>TRP4</i>	Yeast tryptophan biosynthesis	X04273

*Proprietary sequence of Stratagene (La Jolla, California).

†No match in the database; novel EST.

REFERENCES AND NOTES

1. The current EST database (dbEST release 091495) from the National Center for Biotechnology Information (Bethesda, MD) contains a total of 322,225 entries, including 255,645 from the human genome and 21,044 from *Arabidopsis*. Access is available via the World Wide Web (<http://www.ncbi.nlm.nih.gov>).
2. E. M. Meyerowitz and R. E. Pruitt, *Science* 229, 1214 (1985); R. E. Pruitt and E. M. Meyerowitz, *J. Mol. Biol.* 187, 169 (1986); L. Hwang et al., *Plant J.* 1, 367 (1991); P. Jarvis et al., *Plant Mol. Biol.* 24, 685 (1994); L. Le Guan et al., *Mol. Gen. Genet.* 245, 390 (1994).
3. D. Shaon, thesis, Stanford University (1995); and P. O. Brown, in preparation. Microarrays were fabricated on poly-L-lysine-coated microscope slides (Sigma) with a custom-built arraying machine fitted with one printing tip. The tip loaded 1 μ l of PCR product (0.5 mg/ml) from 96-well microtiter plates and deposited \sim 0.005 μ l per slide on 40 slides at a spacing of 500 μ m. The printed slides were rehydrated for 2 hours in a humid chamber, snap-dried at 100°C for 1 min, rinsed in 0.1% SDS, and treated with 0.05% succinic anhydride prepared in buffer consisting of 50% 1-methyl-2-pyrrolidinone and 50% boric acid. The cDNA on the slides was denatured in distilled water for 2 min at 90°C immediately before use. Microarrays were scanned with a laser fluorescent scanner that contained a computer-controlled XY stage and a microscope objective. A mixed gas, multiline laser allowed sequential excitation of the two fluorophores. Emitted light was split according to wavelength and detected with two photomultiplier tubes. Signals were read into a PC with the use of a 12-bit analog-to-digital board. Additional details of microarray fabrication and use may be obtained by means of e-mail (pbrown@cimgm.stanford.edu).
4. F. M. Ausubel et al., Eds., *Current Protocols in Molecular Biology* (Greene & Wiley Interscience, New York, 1994), pp. 4.3.1-4.3.4.
5. Polyadenylated [poly(A)⁺] mRNA was prepared from total RNA with the use of Oligotex-dT resin (Qiagen). Reverse transcription (RT) reactions were carried out with a StrataScript RT-PCR kit (Stratagene) modified as follows: 50- μ l reactions contained 0.1 μ g/ μ l of *Arabidopsis* mRNA, 0.1 ng/ μ l of human AChR mRNA, 0.05 μ g/ μ l of oligo(dT) (21-mer), 1 \times first strand buffer, 0.03 U/ μ l of ribonuclease block, 500 μ M deoxyadenosine triphosphate (dATP), 500 μ M deoxyguanosine triphosphate, 500 μ M dTTP, 40 μ M deoxycytosine triphosphate (dCTP), 40 μ M fluorescein-12-dCTP (or fluorescein-5-dCTP), and 0.03 U/ μ l of StrataScript reverse transcriptase. Reactions were incubated for 60 min at 37°C, precipitated with ethanol, and resuspended in 10 μ l of TE (10 mM Tris-HCl and 1 mM EDTA, pH 8.0). Samples were then heated for 3 min at 94°C and chilled on ice. The RNA was degraded by adding 0.25 μ l of 10 N NaOH followed by a 10-min incubation at 37°C. The samples were neutralized by addition of 2.5 μ l of 1 M Tris-HCl (pH 8.0) and 0.25 μ l of 10 N HCl and precipitated with ethanol. Pellets were washed with 70% ethanol, dried to completion in a speedvac, resuspended in 10 μ l of H₂O, and reduced to 3.0 μ l in a speedvac. Fluorescent nucleotide analogs were obtained from New England Nuclear (DuPont).
6. Hybridization reactions contained 1.0 μ l of fluorescent cDNA synthesis product (5) and 1.0 μ l of hybridization buffer (10 \times saline sodium citrate (SSC) and 0.2% SDS). The 2.0- μ l probe mixtures were aliquoted onto the microarray surface and covered with cover slips (12 mm round). Arrays were transferred to a hybridization chamber (3) and incubated for 18 hours at 65°C. Arrays were washed for 5 min at room temperature (25°C) in low-stringency wash buffer (1 \times SSC and 0.1% SDS), then for 10 min at room temperature in high-stringency wash buffer (0.1 \times SSC and 0.1% SDS). Arrays were scanned in 0.1 \times SSC with the use of a fluorescence laser-scanning device (3).
7. Samples of poly(A)⁺ mRNA (4, 5) were spotted onto nylon membranes (Nyttran) and crosslinked with ultraviolet light with the use of a Stratelink 1800 (Stratagene). Probes were prepared by random priming with the use of a Prime-It II kit (Stratagene) in the presence of [³²P]dATP. Hybridizations were carried out according to the instructions of the manufacturer. Quantitation was performed on a Phosphorimager (Molecular Dynamics).
8. M. Schena and R. W. Davis, *Proc. Natl. Acad. Sci. U.S.A.* 89, 3894 (1992); M. Schena, A. M. Lloyd, R. W. Davis, *Genes Dev.* 7, 367 (1993); M. Schena and R. W. Davis, *Proc. Natl. Acad. Sci. U.S.A.* 91, 8393 (1994).
9. H. Hofte et al., *Plant J.* 4, 1051 (1993); T. Newman et al., *Plant Physiol.* 106, 1241 (1994).
10. N. E. Morton, *Proc. Natl. Acad. Sci. U.S.A.* 88, 7474 (1991); E. D. Green and R. H. Waterston, *J. Am. Med. Assoc.* 266, 1966 (1991); C. Bellanne-Chantelot, *Cell* 70, 1059 (1992); D. R. Cox et al., *Science* 265, 2031 (1994).
11. E. S. Kawasaki et al., *Proc. Natl. Acad. Sci. U.S.A.* 85, 5698 (1988).
12. The laser fluorescent scanner was designed and fabricated in collaboration with S. Smith of Stanford University. Scanner and analysis software was developed by R. X. Xia. The succinic anhydride reaction was suggested by J. Mulligan and J. Van Ness of Darwin Molecular Corporation. Thanks to S. Theologis, C. Somerville, K. Yamamoto, and members of the laboratories of R.W.D. and P.O.B. for critical comments. Supported by the Howard Hughes Medical Institute and by grants from NIH (R21HG00450) (P.O.B.) and R37AG00198 (R.W.D.) and from NSF (MCB9106011) (R.W.D.) and by an NSF graduate fellowship (D.S.). P.O.B. is an assistant investigator of the Howard Hughes Medical Institute.

11 August 1995; accepted 22 September 1995

Gene Therapy in Peripheral Blood Lymphocytes and Bone Marrow for ADA⁻ Immunodeficient Patients

Claudio Bordignon,* Luigi D. Notarangelo, Nadia Nobili, Giuliana Ferrari, Giulia Casorati, Paola Panina, Evelina Mazzolari, Daniela Maggioni, Claudia Rossi, Paolo Servida, Alberto G. Ugazio, Fulvio Mavilio

Adenosine deaminase (ADA) deficiency results in severe combined immunodeficiency, the first genetic disorder treated by gene therapy. Two different retroviral vectors were used to transfer ex vivo the human ADA minigene into bone marrow cells and peripheral blood lymphocytes from two patients undergoing exogenous enzyme replacement therapy. After 2 years of treatment, long-term survival of T and B lymphocytes, marrow cells, and granulocytes expressing the transferred ADA gene was demonstrated and resulted in normalization of the immune repertoire and restoration of cellular and humoral immunity. After discontinuation of treatment, T lymphocytes, derived from transduced peripheral blood lymphocytes, were progressively replaced by marrow-derived T cells in both patients. These results indicate successful gene transfer into long-lasting progenitor cells, producing a functional multilineage progeny.

Severe combined immunodeficiency associated with inherited deficiency of ADA (1) is usually fatal unless affected children are kept in protective isolation or the immune system is reconstituted by bone marrow transplantation from a human leukocyte antigen (HLA)-identical sibling donor (2). This is the therapy of choice, although it is available only for a minority of patients. In recent years, other forms of therapy have been developed, including transplants from haploidentical donors (3, 4), exogenous enzyme replacement (5), and somatic-cell gene therapy (6-9).

We previously reported a preclinical model in which ADA gene transfer and expression

successfully restored immune functions in human ADA-deficient (ADA⁻) peripheral blood lymphocytes (PBLs) in immunodeficient mice in vivo (10, 11). On the basis of these preclinical results, the clinical application of gene therapy for the treatment of ADA⁻ SCID (severe combined immunodeficiency disease) patients who previously failed exogenous enzyme replacement therapy was approved by our Institutional Ethical Committees and by the Italian National Committee for Bioethics (12). In addition to evaluating the safety and efficacy of the gene therapy procedure, the aim of the study was to define the relative role of PBLs and hematopoietic stem cells in the long-term reconstitution of immune functions after retroviral vector-mediated ADA gene transfer. For this purpose, two structurally identical vectors expressing the human ADA complementary DNA (cDNA), distinguishable by the presence of alternative restriction sites in a nonfunctional region of the viral long-terminal repeat (LTR), were used to transduce PBLs and bone marrow (BM) cells independently. This procedure allowed identification of the origin of

C. Bordignon, N. Nobili, G. Ferrari, D. Maggioni, C. Rossi, P. Servida, F. Mavilio, Telethon Gene Therapy Program for Genetic Diseases, DIBIT, Istituto Scientifico H. S. Raffaele, Milan, Italy.
L. D. Notarangelo, E. Mazzolari, A. G. Ugazio, Department of Pediatrics, University of Brescia Medical School, Brescia, Italy.
G. Casorati, Unità di Immunochimica, DIBIT, Istituto Scientifico H. S. Raffaele, Milan, Italy.
P. Panina, Roche Milano Ricerche, Milan, Italy.

*To whom correspondence should be addressed.

Differential gene expression in drug metabolism and toxicology: practicalities, problems and potential

JOHN C. ROCKETT†, DAVID J. ESDAILE‡
and G. GORDON GIBSON*

Molecular Toxicology Laboratory, School of Biological Sciences, University of Surrey,
Guildford, Surrey, GU2 5XH, UK

Received January 8, 1999

1. An important feature of the work of many molecular biologists is identifying which genes are switched on and off in a cell under different environmental conditions or subsequent to xenobiotic challenge. Such information has many uses, including the deciphering of molecular pathways and facilitating the development of new experimental and diagnostic procedures. However, the student of gene hunting should be forgiven for perhaps becoming confused by the mountain of information available as there appears to be almost as many methods of discovering differentially expressed genes as there are research groups using the technique.
2. The aim of this review was to clarify the main methods of differential gene expression analysis and the mechanistic principles underlying them. Also included is a discussion on some of the practical aspects of using this technique. Emphasis is placed on the so-called 'open' systems, which require no prior knowledge of the genes contained within the study model. Whilst these will eventually be replaced by 'closed' systems in the study of human, mouse and other commonly studied laboratory animals, they will remain a powerful tool for those examining less fashionable models.
3. The use of suppression-PCR subtractive hybridization is exemplified in the identification of up- and down-regulated genes in rat liver following exposure to phenobarbital, a well-known inducer of the drug metabolizing enzymes.
4. Differential gene display provides a coherent platform for building libraries and microchip arrays of 'gene fingerprints' characteristic of known enzyme inducers and xenobiotic toxicants, which may be interrogated subsequently for the identification and characterization of xenobiotics of unknown biological properties.

Introduction

It is now apparent that the development of almost all cancers and many non-neoplastic diseases are accompanied by altered gene expression in the affected cells compared to their normal state (Hunter 1991, Wynford-Thomas 1991, Vogelstein and Kinzler 1993, Semenza 1994, Cassidy 1995, Kleinjan and Van Hegningen 1998). Such changes also occur in response to external stimuli such as pathogenic micro-organisms (Rohn *et al.* 1996, Singh *et al.* 1997, Griffin and Krishna 1998, Lunney 1998) and xenobiotics (Sewall *et al.* 1995, Dogra *et al.* 1998, Ramana and Kohli 1998), as well as during the development of undifferentiated cells (Hecht 1998, Rudin and Thompson 1998, Schneider-Maunoury *et al.* 1998). The potential medical and therapeutic benefits of understanding the molecular changes which occur in any given cell in progressing from the normal to the 'altered' state are enormous. Such profiling essentially provides a 'fingerprint' of each step of a

* Author for correspondence; e-mail: g.gibson@surrey.ac.uk

† Current Address: US Environmental Protection Agency, National Health and Environmental Effects, Research Laboratory, Reproductive Toxicology Division, Research Triangle Park, NC 27711, USA.

‡ Rhone-Poulenc Agrochemicals, Toxicology Department, Sophia-Antipolis, Nice, France.

cell's development or response and should help in the elucidation of specific and sensitive biomarkers representing, for example, different types of cancer or previous exposure to certain classes of chemicals that are enzyme inducers.

In drug metabolism, many of the xenobiotic-metabolizing enzymes (including the well-characterized isoforms of cytochrome P450) are inducible by drugs and chemicals in man (Pelkonen *et al.* 1998), predominantly involving transcriptional activation of not only the cognate cytochrome P450 genes, but additional cellular proteins which may be crucial to the phenomenon of induction. Accordingly, the development of methodology to identify and assess the full complement of genes that are either up- or down-regulated by inducers are crucial in the development of knowledge to understand the precise molecular mechanisms of enzyme induction and how this relates to drug action. Similarly, in the field of chemical-induced toxicity, it is now becoming increasingly obvious that most adverse reactions to drugs and chemicals are the result of multiple gene regulation, some of which are causal and some of which are casually-related to the toxicological phenomenon *per se*. This observation has led to an upsurge in interest in gene-profiling technologies which differentiate between the control and toxin-treated gene pools in target tissues and is, therefore, of value in rationalizing the molecular mechanisms of xenobiotic-induced toxicity. Knowledge of toxin-dependent gene regulation in target tissues is not solely an academic pursuit as much interest has been generated in the pharmaceutical industry to harness this technology in the early identification of toxic drug candidates, thereby shortening the developmental process and contributing substantially to the safety assessment of new drugs. For example, if the gene profile in response to say a testicular toxin that has been well-characterized *in vivo* could be determined in the testis, then this profile would be representative of all new drug candidates which act via this specific molecular mechanism of toxicity, thereby providing a useful and coherent approach to the early detection of such toxicants. Whereas it would be informative to know the identity and functionality of all genes up/down regulated by such toxicants, this would appear a longer term goal, as the majority of human genes have not yet been sequenced, far less their functionality determined. However, the current use of gene profiling yields a *pattern* of gene changes for a xenobiotic of unknown toxicity which may be matched to that of well-characterized toxins, thus alerting the toxicologist to possible *in vivo* similarities between the unknown and the standard, thereby providing a platform for more extensive toxicological examination. Such approaches are beginning to gain momentum, in that several biotechnology companies are commercially producing 'gene chips' or 'gene arrays' that may be interrogated for toxicity assessment of xenobiotics. These chips consist of hundreds/thousands of genes, some of which are degenerate in the sense that not all of the genes are mechanistically-related to any one toxicological phenomenon. Whereas these chips are useful in broad-spectrum screening, they are maturing at a substantial rate, in that gene arrays are now becoming more specific, e.g. chips for the identification of changes in growth factor families that contribute to the aetiology and development of chemically-induced neoplasias.

Although documenting and explaining these genetic changes presents a formidable obstacle to understanding the different mechanisms of development and disease progression, the technology is now available to begin attempting this difficult challenge. Indeed, several 'differential expression analysis' methods have been developed which facilitate the identification of gene products that demonstrate

altered expression in cells of one population compared to another. These methods have been used to identify differential gene expression in many situations, including invading pathogenic microbes (Zhao *et al.* 1998), in cells responding to extracellular and intracellular microbial invasion (Duguid and Dinauer 1990, Ragno *et al.* 1997, Maldarelli *et al.* 1998), in chemically treated cells (Syed *et al.* 1997, Rockett *et al.* 1999), neoplastic cells (Liang *et al.* 1992, Chang and Terzaghi-Howe 1998), activated cells (Gurskaya *et al.* 1996, Wan *et al.* 1996), differentiated cells (Hara *et al.* 1991, Guimaraes *et al.* 1995a, b), and different cell types (Davis *et al.* 1984, Hedrick *et al.* 1984, Xhu *et al.* 1998). Although differential expression analysis technologies are applicable to a broad range of models, perhaps their most important advantage is that, in most cases, absolutely no prior knowledge of the specific genes which are up- or down-regulated is required.

The field of differential expression analysis is a large and complex one, with many techniques available to the potential user. These can be categorized into several methodological approaches, including:

- (1) Differential screening,
- (2) Subtractive hybridization (SH) (includes methods such as chemical cross-linking subtraction—CCLS, suppression-PCR subtractive hybridization—SSH, and representational difference analysis—RDA),
- (3) Differential display (DD),
- (4) Restriction endonuclease facilitated analysis (including serial analysis of gene expression—SAGE—and gene expression fingerprinting—GEF),
- (5) Gene expression arrays, and
- (6) Expressed sequence tag (EST) analysis.

The above approaches have been used successfully to isolate differentially expressed genes in different model systems. However, each method has its own subtle (and sometimes not so subtle) characteristics which incur various advantages and disadvantages. Accordingly, it is the purpose of this review to clarify the mechanistic principles underlying the main differential expression methods and to highlight some of the broader considerations and implications of this very powerful and increasingly popular technique. Specifically, we will concentrate on the so-called 'open' systems, namely those which do not require any knowledge of gene sequences and, therefore, are useful for isolating unknown genes. Two 'closed' systems (those utilising previously identified gene sequences), EST analysis and the use of DNA arrays, will also be considered briefly for completeness. Whilst emphasis will often be placed on suppression PCR subtractive hybridization (SSH, the approach employed in this laboratory), it is the aim of the authors to highlight, wherever possible, those areas of common interest to those who use, or intend to use, differential gene expression analysis.

Differential cDNA library screening (DS)

Despite the development of multiple technological advances which have recently brought the field of gene expression profiling to the forefront of molecular analysis, recognition of the importance of differential gene expression and characterization of differentially expressed genes has existed for many years. One of the original approaches used to identify such genes was described 20 years ago by St John and Davis (1979). These authors developed a method, termed 'differential plaque filter

hybridization', which was used to isolate galactose-inducible DNA sequences from yeast. The theory is simple: a genomic DNA library is prepared from normal, unstimulated cells of the test organism/tissue and multiple filter replicas are prepared. These replica blots are probed with radioactively (or otherwise) labelled complex cDNA probes prepared from the control and test cell mRNA populations. Those mRNAs which are differentially expressed in the treated cell population will show a positive signal only on the filter probed with cDNA from the treated cells. Furthermore, labelled cDNA from different test conditions can be used to probe multiple blots, thereby enabling the identification of mRNAs which are only up-regulated under certain conditions. For example, St John and Davis (1979) screened replica filters with acetate-, glucose- and galactose-derived probes in order to obtain genes induced specifically by galactose metabolism. Although groundbreaking in its time this method is now considered insensitive and time-consuming, as up to 2 months are required to complete the identification of genes which are differentially expressed in the test population. In addition, there is no convenient way to check that the procedure has worked until the whole process has been completed.

Subtractive Hybridization (SH)

The developing concept of differential gene expression and the success of early approaches such as that described by St John and Davis (1979) soon gave rise to a search for more convenient methods of analysis. One of the first to be developed was SH, numerous variations of which have since been reported (see below). In general, this approach involves hybridization of mRNA/cDNA from one population (tester) to excess mRNA/cDNA from another (driver), followed by separation of the unhybridized tester fraction (differentially expressed) from the hybridized common sequences. This step has been achieved physically, chemically and through the use of selective polymerase chain reaction (PCR) techniques.

Physical separation

Original subtractive hybridization technology involved the physical separation of hybridized common species from unique single stranded species. Several methods of achieving this have been described, including hydroxyapatite chromatography (Sargent and Dawid 1983), avidin-biotin technology (Duguid and Dinauer 1990) and oligodT-latex separation (Hara *et al.* 1991). In the first approach, common mRNA species are removed by cDNA (from test cells)-mRNA (from control cells) subtractive hybridization followed by hydroxyapatite chromatography, as hydroxyapatite specifically adsorbs the cDNA-mRNA hybrids. The unabsorbed cDNA is then used either for the construction of a cDNA library of differentially expressed genes (Sargent and Dawid 1983, Schneider *et al.* 1988) or directly as a probe to screen a preselected library (Zimmerman *et al.* 1980, Davis *et al.* 1984, Hedrick *et al.* 1984). A schematic diagram of the procedure is shown in figure 1.

Less rigorous physical separation procedures coupled with sensitivity enhancing PCR steps were later developed as a means to overcome some of the problems encountered with the hydroxyapatite procedure. For example, Duguid and Dinauer (1990) described a method of subtraction utilizing biotin-affinity systems as a means to remove hybridized common sequences. In this process, both the control and tester mRNA populations are first converted to cDNA and an adaptor ('oligovector',

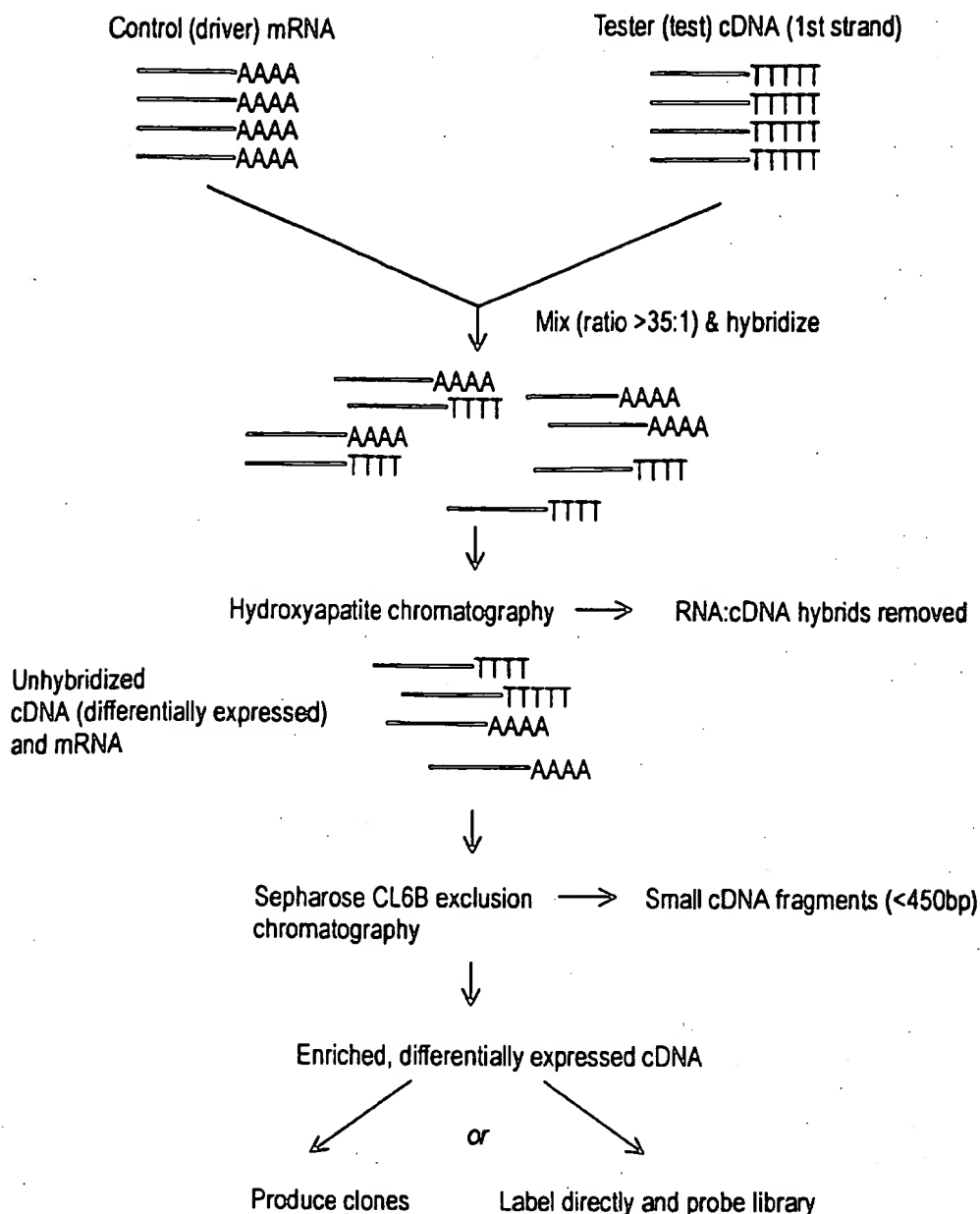


Figure 1. The hydroxyapatite method of subtractive hybridization. cDNA derived from the treated/alterd (tester) population is mixed with a large excess of mRNA from the control (driver) population. Following hybridization, mRNA-cDNA hybrids are removed by hydroxyapatite chromatography. The only cDNAs which remain are those which are differentially expressed in the treated/alterd population. In order to facilitate the recovery of full length clones, small cDNA fragments are removed by exclusion chromatography. The remaining cDNAs are then cloned into a vector for sequencing, or labelled and used directly to probe a library, as described by Sargent and Dawid (1983).

containing a restriction site) ligated to both sides. Both populations are then amplified by PCR, but the driver cDNA population is subsequently digested with the adaptor-containing restriction endonuclease. This serves to cleave the oligo-vector and reduce the amplification potential of the control population. The digested control population is then biotinylated and an excess mixed with tester cDNA. Following denaturation and hybridization, the mix is applied to a biocytin column (streptavidin may also be used) to remove the control population, including heteroduplexes formed by annealing of common sequences from the tester population. The procedure is repeated several times following the addition of fresh

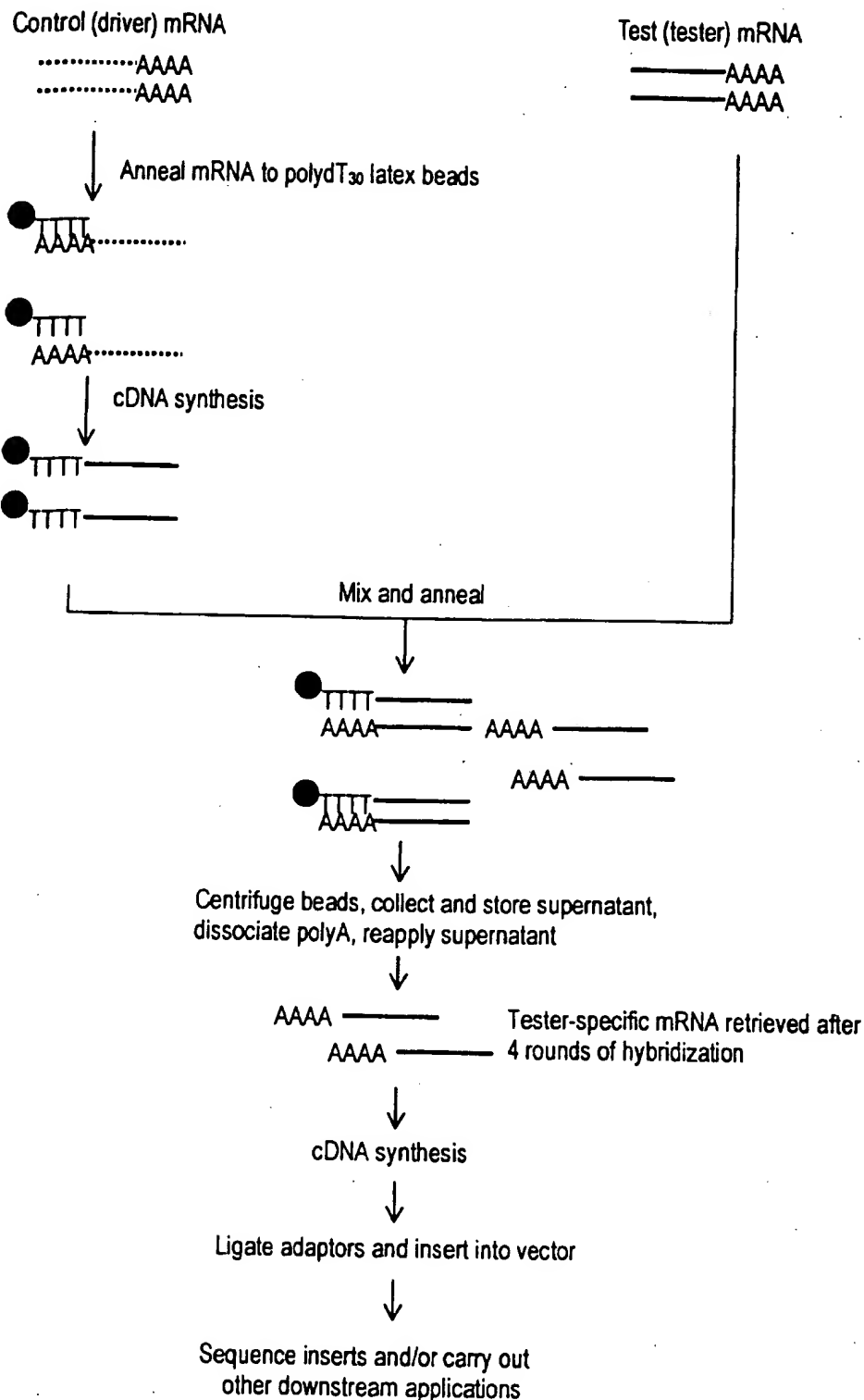


Figure 2. The use of oligodT₃₀ latex to perform subtractive hybridization. mRNA extracted from the control (driver) population is converted to anchored cDNA using polydT oligonucleotides attached to latex beads. mRNA from the treated/alterd (tester) population is repeatedly hybridized against an excess of the anchored driver cDNA. The final population of mRNA is tester specific and can be converted into cDNA for cloning and other downstream applications, as described by Hara *et al.* (1991).

control cDNA. In order to further enrich those species differentially expressed in the tester cDNA, the subtracted tester population is amplified by PCR following every second subtraction cycle. After six cycles of subtraction (three reamplification steps) the reaction mix is ligated into a vector for further analysis.

In a slightly different approach, Hara *et al.* (1991) utilized a method whereby oligo(dT₃₀) primers attached to a latex substrate are used to first capture mRNA extracted from the control population. Following 1st strand cDNA synthesis, the RNA strand of the heteroduplexes is removed by heat denaturation and centrifugation (the cDNA-oligotex-dT₃₀ forms a pellet and the supernatant is removed). A quantity of tester mRNA is then repeatedly hybridized to the immobilized control (driver) cDNA (which is present in 20-fold excess). After several rounds of hybridization the only mRNA molecules left in the tester mRNA population are those which are not found in the driver cDNA-oligotex-dT₃₀ population. These tester-specific mRNA species are then converted to cDNA and, following the addition of adaptor sequences, amplified by PCR. The PCR products are then ligated into a vector for further analysis using restriction sites incorporated into the PCR primers. A schematic illustration of this subtraction process is shown in figure 2.

However, all these methods utilising physical separation have been described as inefficient due to the requirement for large starting amounts of mRNA, significant loss of material during the separation process and a need for several rounds of hybridization. Hence, new methods of differential expression analysis have recently been designed to eliminate these problems.

Chemical Cross-Linking Subtraction (CCLS)

In this technique, originally described by Hampson *et al.* (1992), driver mRNA is mixed with tester cDNA (1st strand only) in a ratio of > 20:1. The common sequences form cDNA:mRNA hybrids, leaving the tester specific species as single stranded cDNA. Instead of physically separating these hybrids, they are inactivated chemically using 2,5 diaziridinyl-1,4-benzoquinone (DZQ). Labelled probes are then synthesized from the remaining single stranded cDNA species (unreacted mRNA species remaining from the driver are not converted into probe material due to specificity of Sequenase T7 DNA polymerase used to make the probe) and used to screen a cDNA library made from the tester cell population. A schematic diagram of the system is shown in figure 3.

It has been shown that the differentially expressed sequences can be enriched at least 300-fold with one round of subtraction (Hampson *et al.* 1992), and that the technique should allow isolation of cDNAs derived from transcripts that are present at less than 50 copies per cell. This equates to genes at the low end of intermediate abundance (see table 1). The main advantages of the CCLS approach are that it is rapid, technically simple and also produces fewer false positives than other differential expression analysis methods. However, like the physical separation protocols, a major drawback with CCLS is the large amount of starting material required (at least 10 µg RNA). Consequently, the technique has recently been refined so that a renewable source of RNA can be generated. The degenerate random oligonucleotide primed (DROP) adaptation (Hampson *et al.* 1996, Hampson and Hampson 1997) uses random hexanucleotide sequences to prime solid phase-synthesized cDNA. Since each primer includes a T7 polymerase promotor sequence

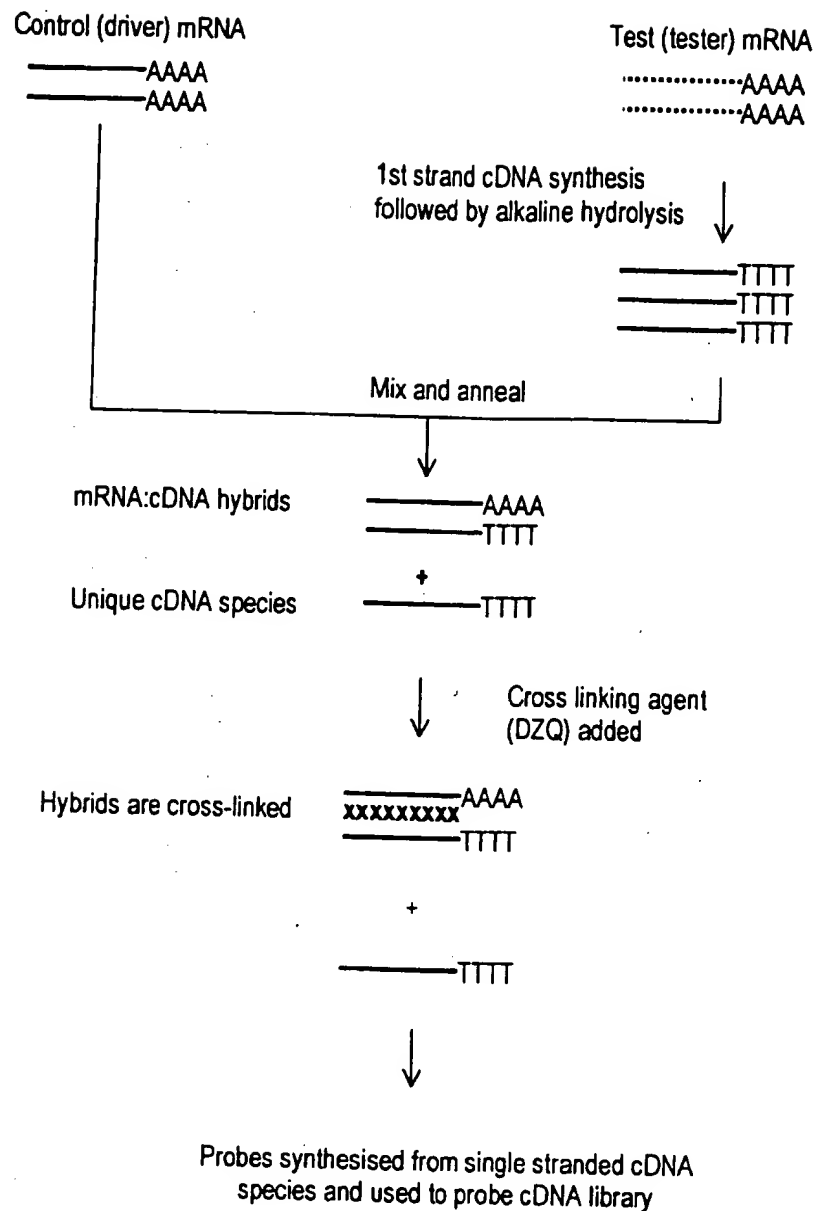


Figure 3. Chemical cross-linking subtraction. Excess driver mRNA is mixed with 1st strand tester cDNA. The common sequences form mRNA:cDNA hybrids which are cross linked with 2,5 diaziridinyl-1,4-benzoquinone (DZQ) and the remaining cDNA sequences are differentially expressed in the tester population. Probes are made from these sequences using Sequenase 2.0 DNA polymerase, which lacks reverse transcriptase activity and, therefore, does not react with the remaining mRNA molecules from the driver. The labelled probes are then used to screen a cDNA library for clones of differentially expressed sequences. Adapted from Walter *et al.* (1996), with permission.

Table 1. The abundance of mRNA species and classes in a typical mammalian cell.

mRNA class	Copies of each species/cell	No. of mRNA species in class	Mean % of each species in class	Mean mass (ng) of each species/ μ g total RNA
Abundant	12 000	4	3.3	1.65
Intermediate	300	500	0.08	0.04
Rare	15	11 000	0.004	0.002

Modified from Bertoli *et al.* (1995).

at the 5' end, the final pool of random cDNA fragments is a PCR-renewable cDNA population which is representative of the expressed gene pool and can be used to synthesize sense RNA for use as driver material. Furthermore, if the final pool of random cDNA fragments is reamplified using biotinylated T7 primer and random hexamer, the product can be captured with streptavidin beads and the antisense strand eluted for use as tester. Since both target and driver can be generated from the same DROP product, subtraction can be performed in both directions (i.e. for up- and down-regulated species) between two different DROP products.

Representational Difference Analysis (RDA)

RDA of cDNA (Hubank and Schatz 1994) is an extension of the technique originally applied to genomic DNA as a means of identifying differences between two complex genomes (Lisitsyn *et al.* 1993). It is a process of subtraction and amplification involving subtractive hybridization of the tester in the presence of excess driver. Sequences in the tester that have homologues in the driver are rendered unamplifiable, whereas those genes expressed only in the tester retain the ability to be amplified by PCR. The procedure is shown schematically in figure 4.

In essence, the driver and tester mRNA populations are first converted to cDNA and amplified by PCR following the ligation of an adaptor. The adaptors are then removed from both populations and a new (different) adaptor ligated to the amplified tester population only. Driver and tester populations are next melted and hybridized together in a ratio of 100:1. Following hybridization, only tester:tester homohybrids have 5' adaptors at each end of the DNA duplex and can, thus, be filled in at both 3' ends. Hence, only these molecules are amplified exponentially during the subsequent PCR step. Although tester:driver heterohybrids are present, they only amplify in a linear fashion, since the strand derived from the driver has no adaptor to which the primer can bind. Driver:driver heterohybrids have no adaptors and, therefore, are not amplified. Single stranded molecules are digested with mung bean nuclease before a further PCR-enrichment of the tester:tester homohybrids. The adaptors on the amplified tester population are then replaced and the whole process repeated a further two or three times using an increasing excess of driver (Hubank and Schatz used a tester:driver ratio of 1:400, 1:80000 and 1:800000 for the second, third and fourth hybridizations, respectively). Different adaptors are ligated to the tester between successive rounds of hybridization and amplification to prevent the accumulation of PCR products that might interfere with subsequent amplifications. The final display is a series of differentially expressed gene products easily observable on an ethidium bromide gel.

The main advantages of RDA are that it offers a reproducible and sensitive approach to the analysis of differentially expressed genes. Hubank and Schatz (1994) reported that they were able to isolate genes that were differentially expressed in substantially less than 1% of the cells from which the tester is derived. Perhaps the main drawback is that multiple rounds of ligation, hybridization, amplification and digestion are required. The procedure is, therefore, lengthier than many other differential display approaches and provides more opportunity for operator-induced error to occur. Although the generation of false positives has been noted, this has been solved to some degree by O'Neill and Sinclair (1997) through the use of HPLC-purified adaptors. These are free of the truncated adaptors which appear to be a major source of the false positive bands. A very similar technique to RDA, termed linker capture subtraction (LCS) was described by Yang and Sytowski (1996).

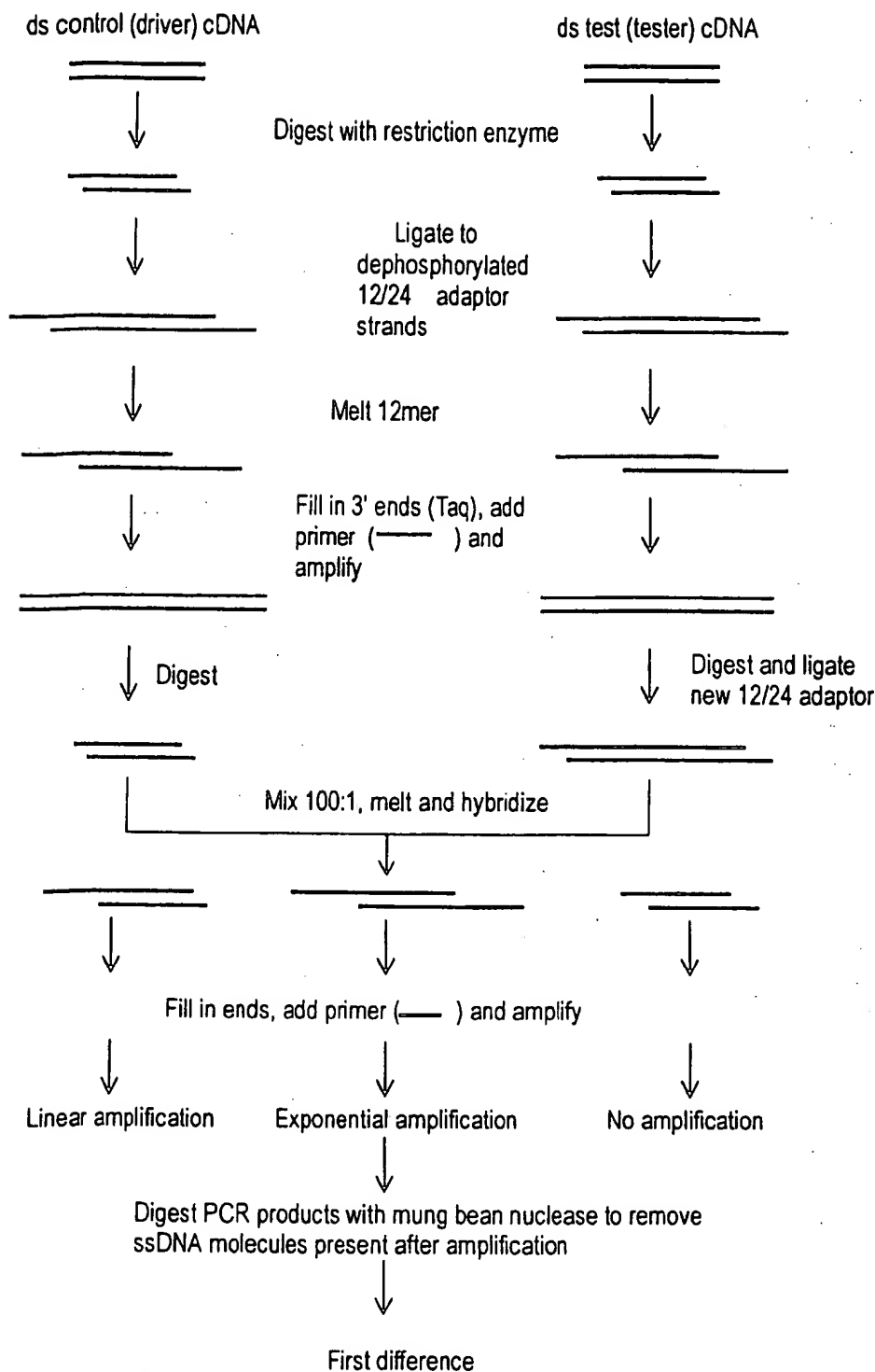


Figure 4. The representational difference analysis (RDA) technique. Driver and tester cDNA are digested with a 4-cutter restriction enzyme such as *DpnII*. The 1st set of 12/24 adaptor strands (oligonucleotides) are ligated to each other and the digested cDNA products. The 12mer is subsequently melted away and the 3' ends filled in using Taq DNA polymerase. Each cDNA population is then amplified using PCR, following which the 1st set of adaptors is removed with *DpnII*. A second set of 12/24 adaptor strands is then added to the amplified tester cDNA population, after which the tester is hybridized against a large excess of driver. The 12mer adaptors are melted and the 3' ends filled in as before. PCR is carried out with primers identical to the new 24mer adaptor. Thus, the only hybridization products which are exponentially amplified are those which are tester:tester combinations. Following PCR, ssDNA products are removed with mung bean nuclease, leaving the 'first difference product'. This is digested and a third set of 12/24 adaptors added before repeating the subtraction process from the hybridization stage. The process is repeated to the 3rd or 4th difference product, as described by Lisitsyn *et al.* (1993) and Hubank and Schatz (1994).

Suppression PCR Subtractive Hybridization (SSH)

The most recent adaptation of the SH approach to differential expression analysis was first described by Diatchenko *et al.* (1996) and Gurskaya *et al.* (1996). They reported that a 1000–5000 fold enrichment of rare cDNAs (equivalent to isolating mRNAs present at only a few copies per cell) can be obtained without the need for multiple hybridizations/subtractions. Instead of physical or chemical removal of the common sequences, a PCR-based suppression system is used (see figure 5).

In SSH, excess driver cDNA is added to two portions of the tester cDNA which have been ligated with different adaptors. A first round of hybridization serves to enrich differentially expressed genes and equalize rare and abundant messages. Equalization occurs since reannealing is more rapid for abundant molecules than for rarer molecules due to the second order kinetics of hybridization (James and Higgins 1985). The two primary hybridization mixes are then mixed together in the presence of excess driver and allowed to hybridize further. This step permits the annealing of single stranded complementary sequences which did not hybridize in the primary hybridization, and in doing so generates templates for PCR amplification. Although there are several possible combinations of the single stranded molecules present in the secondary hybridization mix, only one particular combination (differentially expressed in the tester cDNA composed of complimentary strands having different adaptors) can amplify exponentially.

Having obtained the final differential display, two options are available if cloning of cDNAs is desired. One is to transform the whole of the final PCR reaction into competent cells. Transformed colonies can then be isolated and their inserts characterized by sequencing, restriction analysis or PCR. Alternatively, the final PCR products can be resolved on a gel and the individual bands excised, reamplified and cloned. The first approach is technically simpler and less time consuming. However, ligation/transformation reactions are known to be biased towards the cloning of smaller molecules, and so the final population of clones will probably not contain a representative selection of the larger products. In addition, although equalization theoretically occurs, observations in this laboratory suggest that this is by no means perfectly accomplished. Consequently, some gene species are present in a higher number than others and this will be represented in the final population of clones. Thus, in order to obtain a substantial proportion of those gene species that actually demonstrate differential expression in the tester population, the number of clones that will have to be screened after this step may be substantial. The second approach is initially more time consuming and technically demanding. However, it would appear to offer better prospects for cloning larger and low abundance gel products. In addition, one can incorporate a screening step that differentiates different products of different sequences but of the same size (HA-staining, see later). In this way, a good idea of the final number of clones to be isolated and identified can be achieved.

An alternative (or even complementary) approach is to use the final differential display reaction to screen a cDNA library to isolate full length clones for further characterization, or a DNA array (see later) to quickly identify known genes. SSH has been used in this laboratory to begin characterization of the short-term gene expression profiles of enzyme-inducers such as phenobarbital (Rockett *et al.* 1997) and Wy-14,643 (Rockett *et al.* unpublished observations). The isolation of differentially expressed genes in this manner enables the construction of a fingerprint

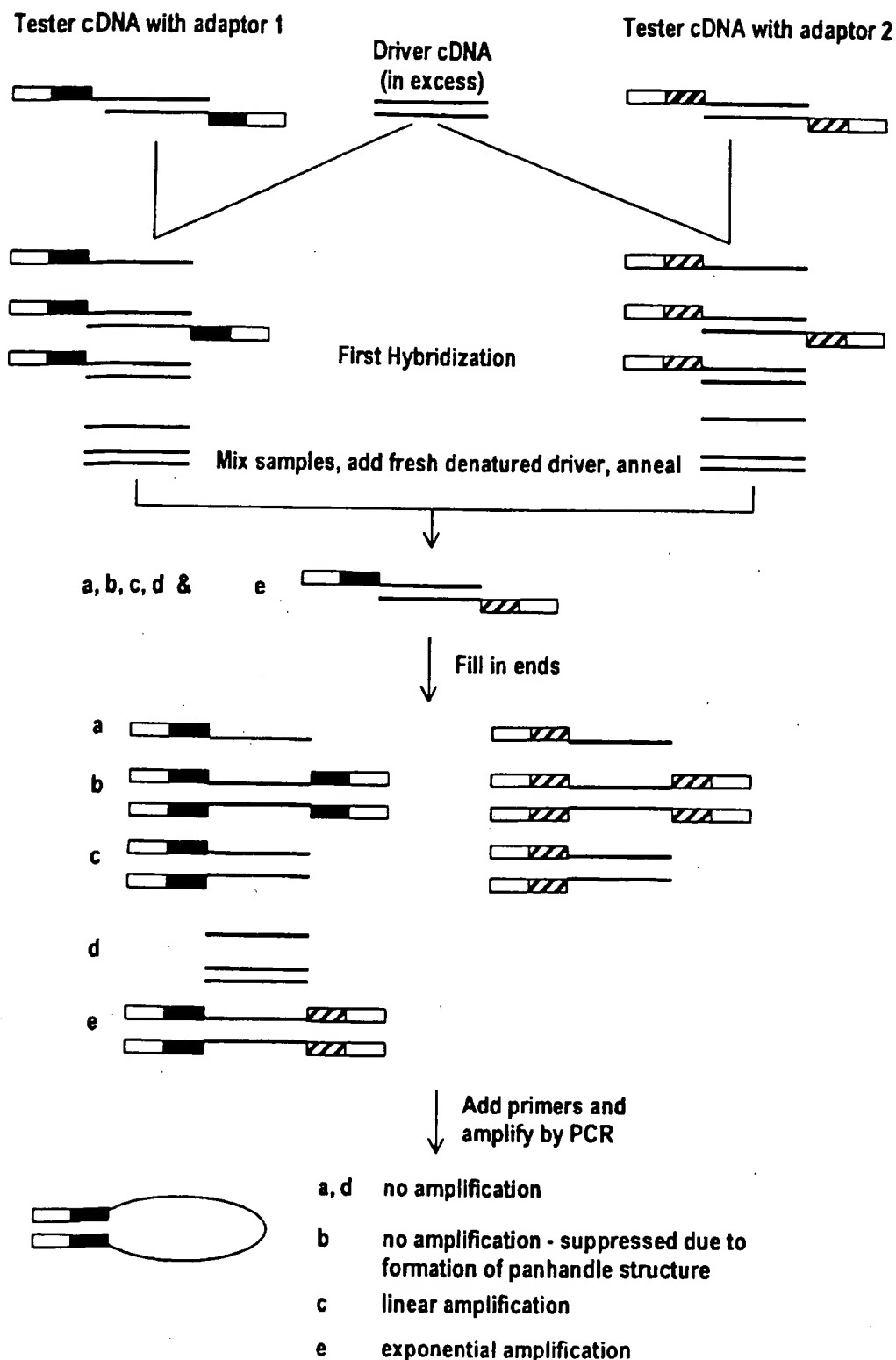


Figure 5. PCR-select cDNA subtraction. In the primary hybridization, an excess of driver cDNA is added to each tester cDNA population. The samples are heat denatured and allowed to hybridize for between 3 and 8 h. This serves two purposes: (1) to equalize rare and abundant molecules; and (2) to enrich for differentially expressed sequences—cDNAs that are not differentially expressed form type c molecules with the driver. In the secondary hybridization, the two primary hybridizations are mixed together without denaturing. Fresh denatured driver can also be added at this point to allow further enrichment of differentially expressed sequences. Type e molecules are formed in this secondary hybridization which are subsequently amplified using two rounds of PCR. The final products can be visualized on an agarose gel, labelled directly or cloned into a vector for downstream manipulation. As described by Diatchenko *et al.* (1996) and Gurskaya *et al.* (1996), with permission.

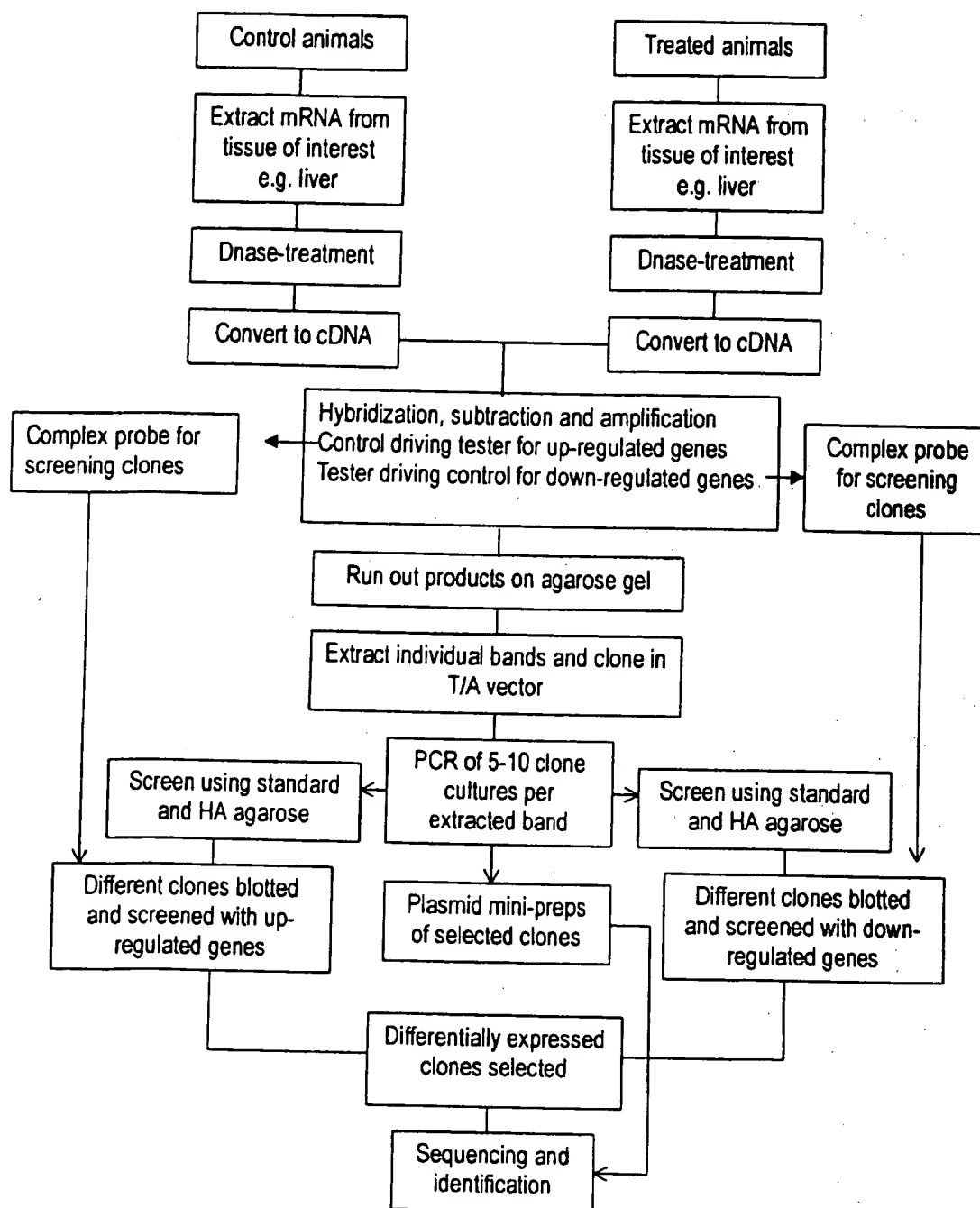


Figure 6. Flow diagram showing method used in this laboratory to isolate and identify clones of genes which are differentially expressed in rat liver following short term exposure to the enzyme inducers, phenobarbital and Wy-14,643.

of expressed genes which are unique to each compound and time/dose point. Such information could be useful in short-term characterization of the toxic potential of new compounds by comparing the gene-expression profiles they elicit with those produced by known inducers. Figure 6 shows a flow diagram of the method used to isolate, verify and clone differentially expressed genes, and figure 7 shows expression profiles obtained from a typical SSH experiment. Subsequent sub-cloning of the individual bands, sequencing and gene data base interrogation reveals many genes which are either up- or down-regulated by phenobarbital in the rat (tables 2 and 3).

One of the advantages in using the SSH approach is that no prior knowledge is required of which specific genes are up/down-regulated subsequent to xenobiotic

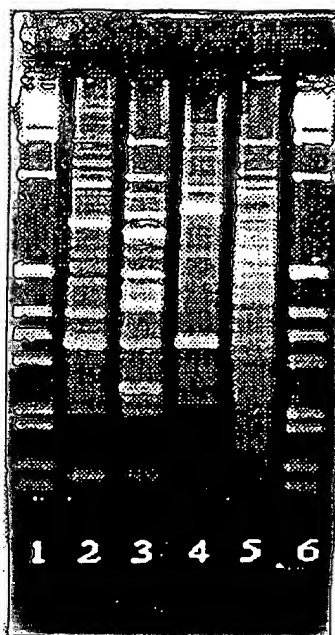


Figure 7. SSH display patterns obtained from rat liver following 3-day treatment with WY-14,643 or phenobarbital. mRNA extracted from control and treated livers was used to generate the differential displays using the PCR-Select cDNA subtraction kit (Clontech). Lane: 1—1kb ladder; 2—genes upregulated following Wy,14-643 treatment; 3—genes downregulated following Wy,14-643 treatment; 4—genes upregulated following phenobarbital treatment; 5—genes downregulated following phenobarbital treatment; 6—1kb ladder. Reproduced from Rockett *et al.* (1997), with permission.

exposure, and an almost complete complement of genes are obtained. For example, the peroxisome proliferator and non-genotoxic hepatocarcinogen Wy,14,643, up-regulates at least 28 genes and down-regulates at least 15 in the rat (a sensitive species) and produces 48 up- and 37 down-regulated genes in the guinea pig, a resistant species (Rockett, Swales, Esda and Gibson, unpublished observations). One of these genes, CD81, was up-regulated in the rat and down-regulated in the guinea pig following Wy-14,643 treatment. CD81 (alternatively named TAPA-1) is a widely expressed cell surface protein which is involved in a large number of cellular processes including adhesion, activation, proliferation and differentiation (Levy *et al.* 1998). Since all of these functions are altered to some extent in the phenomena of hepatomegaly and non-genotoxic hepatocarcinogenesis, it is intriguing, and probably mechanistically-relevant, that CD81 expression is differentially regulated in a resistant and susceptible species. However, the down-side of this approach is that the majority of genes can be sequenced and matched to database sequences, but the latter are predominantly expressed sequence tags or genes of completely unknown function, thus partially obscuring a realistic overall assessment of the critical genes of genuine biological interest. Notwithstanding the lack of complete functional identification of altered gene expression, such gene profiling studies essentially provides a 'molecular fingerprint' in response to xenobiotic challenge, thereby serving as a mechanistically-relevant platform for further detailed investigations.

Differential Display (DD)

Originally described as 'RNA fingerprinting by arbitrarily primed PCR' (Liang and Pardee 1992) this method is now more commonly referred to as 'differential

Table 2. Genes up-regulated in rat liver following 3-day exposure to phenobarbital.

Band number (approximate size in bp)	Highest sequence similarity	FASTA-EMBL gene identification
5 (1300)	93.5%	CYP2B1
7 (1000)	95.1%	Preproalbumin Serum albumin mRNA
8 (950)	98.3%	NCI-CGAP-Pr1 <i>H. sapiens</i> (EST)
10 (850)	95.7%	CYP2B1
11 (800)	Clone 1 94.9%	CYP2B1
	Clone 2 75.3%	CYP2B2
12 (750)	93.8%	TRPM-2 mRNA Sulfated glycoprotein
15 (600)	92.9%	Preproalbumin Serum albumin mRNA
16 (55)	Clone 1 95.2%	CYP2B1
	Clone 2 93.6%	Haptoglobin mRNA partial alpha
21 (350)	99.3%	18S, 5.8S & 28S rRNA

Bands 1–4, 6, 9, 13, 14, and 17–20 are shown to be false positives by dot blot analysis and, therefore, are not sequenced. Derived from Rockett *et al.* (1997). It should be noted that the above genes do not represent the complete spectrum of genes which are up-regulated in rat liver by phenobarbital, but simply represents the genes sequenced and identified to date.

Table 3. Genes down-regulated in rat liver following 3-day exposure to phenobarbital.

Band number (approximate size in bp)	Highest sequence similarity	FASTA-EMBL gene identification
1 (1500)	95.3%	3-oxoacyl-CoA thiolase
2 (1200)	92.3%	Hemopoxin mRNA
3 (1000)	91.7%	Alpha-2u-globulin mRNA
7 (700)	Clone 1 77.2%	<i>M. musculus</i> C1 inhibitor
	Clone 2 94.5%	Electron transfer flavoprotein
	Clone 3 91.0%	<i>M. musculus</i> Topoisomerase 1 (Topo 1)
8 (650)	Clone 1 86.9%	Soares 2NbMT <i>M. musculus</i> (EST)
	Clone 2 96.2%	Alpha-2u-globulin (s-type) mRNA
9 (600)	Clone 1 86.9%	Soares mouse NML <i>M. musculus</i> (EST)
	Clone 2 82.0%	Soares p3NMF 19.5 <i>M. musculus</i> (EST)
10 (550)	73.8%	Soares mouse NML <i>M. musculus</i> (EST)
11 (525)	95.7%	NCI-CGAP-Pr1 <i>H. sapiens</i> (EST)
12 (375)	100.0%	Ribosomal protein
13 (23)	Clone 1 97.2%	Soares mouse embryo NbME135 (EST)
	Clone 2 100.0%	Fibrinogen B-beta-chain
	Clone 3 100.0%	Apolipoprotein E gene
14 (170)	96.0%	Soares p3NMF19.5 <i>M. musculus</i> (EST)
15 (140)	97.3%	Stratagene mouse testis (EST)
Others: (300)	96.7%	<i>R. norvegicus</i> RASP 1 mRNA
(275)	93.1%	Soares mouse mammary gland (EST)

EST = Expressed sequence tag. Bands 4–6 were shown to be false positives by dot blot analysis and, therefore, were not sequenced. Derived from Rockett *et al.* (1997). It should be noted that the above genes do not represent the complete spectrum of genes which are down-regulated in rat liver by phenobarbital, but simply represents the genes sequenced and identified to date.

display' (DD). In this method, all the mRNA species in the control and treated cell populations are amplified in separate reactions using reverse transcriptase-PCR (RT-PCR). The products are then run side-by-side on sequencing gels. Those bands which are present in one display only, or which are much more intense in one

display compared to the other, are differentially expressed and may be recovered for further characterization. One advantage of this system is the speed with which it can be carried out—2 days to obtain a display and as little as a week to make and identify clones.

Two commonly used variations are based on different methods of priming the reverse transcription step (figure 8). One is to use an oligo dT with a 2-base 'anchor' at the 3'-end, e.g. 5' (dT_n)CA 3' (Liang and Pardee 1992). Alternatively, an arbitrary primer may be used for 1st strand cDNA synthesis (Welsh *et al.* 1992). This variant of RNA fingerprinting has also been called 'RAP' (RNA Arbitrarily Primed)-PCR. One advantage of this second approach is that PCR products may be derived from anywhere in the RNA, including open reading frames. In addition, it can be used for mRNAs that are not polyadenylated, such as many bacterial mRNAs (Wong and McClelland 1994). In both cases, following reverse transcription and denaturation, second strand cDNA synthesis is carried out with an arbitrary primer (*arbitrary* primers have a single base at each position, as compared to *random* primers, which contain a mixture of all four bases at each position). The resulting PCR, thus, produces a series of products which, depending on the system (primer length and composition, polymerase and gel system), usually includes 50–100 products per primer set (Band and Sager 1989). When a combination of different dT-anchors and arbitrary primers are used, almost all mRNA species from a cell can be amplified. When the cDNA products from two different populations are analysed side by side on a polyacrylamide gel, differences in expression can be identified and the appropriate bands recovered for cloning and further analysis.

Although DD is perhaps the most popular approach used today for identifying differentially expressed genes, it does suffer from several perceived disadvantages:

- (1) It may have a strong bias towards high copy number mRNAs (Bertioli *et al.* 1995), although this has been disputed (Wan *et al.* 1996) and the isolation of very low abundance genes may be achieved in certain circumstances (Guimeraes *et al.* 1995a).
- (2) The cDNAs obtained often only represent the extreme 3' end of the mRNA (often the 3'-untranslated region), although this may not always be the case (Guimeraes *et al.* 1995a). Since the 3' end is often not included in Genbank and shows variation between organisms, cDNAs identified by DD cannot always be matched with their genes, even if they have been identified.
- (3) The pattern of differential expression seen on the display often cannot be reproduced on Northern blots, with false positives arising in up to 70% of cases (Sun *et al.* 1994). Some adaptations have been shown to reduce false positives, including the use of two reverse transcriptases (Sung and Denman 1997), comparison of uninduced and induced cells over a time course (Burn *et al.* 1994) and comparison of DDPCR-products from two uninduced and two induced lines (Sompayrac *et al.* 1995). The latter authors also reported that the use of cytoplasmic RNA rather than total RNA reduces false positives arising from nuclear RNA that is not transported to the cytoplasm.

Further details of the background, strengths and weaknesses of the DD technique can be obtained from a review by McClelland *et al.* (1996) and from articles by Liang *et al.* (1995) and Wan *et al.* (1996).

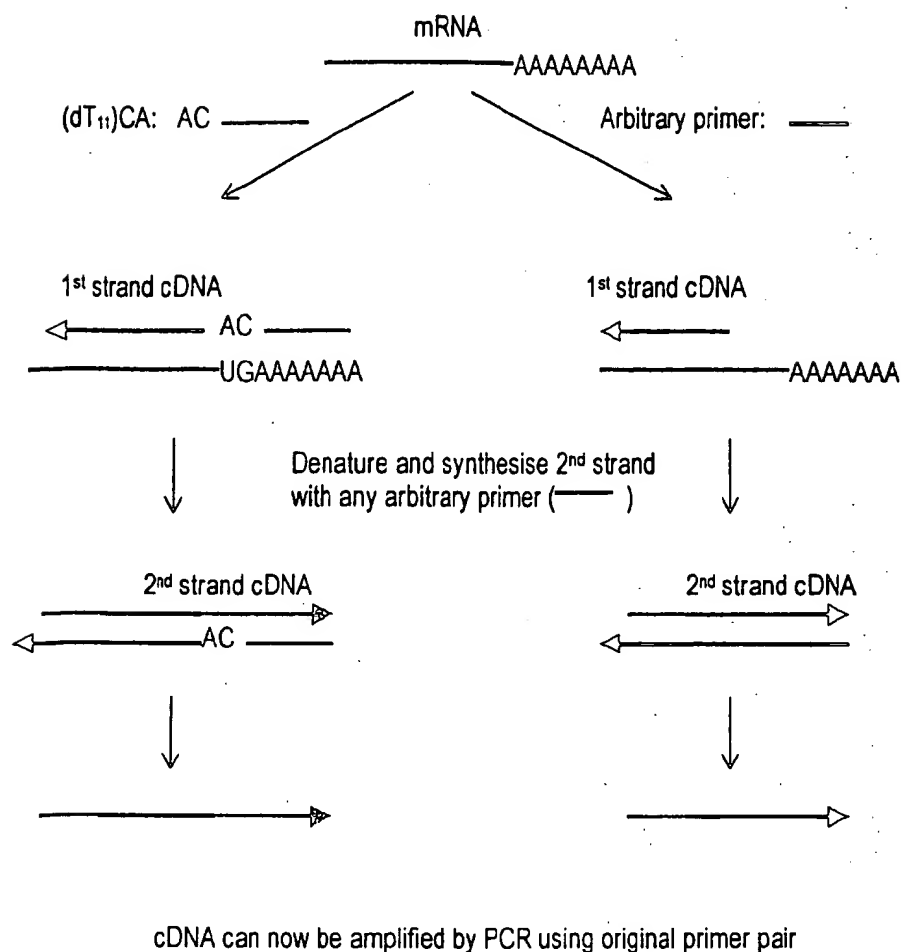


Figure 8. Two approaches to differential display (DD) analysis. 1st strand synthesis can be carried out either with a polydT₁₁NN primer (where N = G, C or A) or with an arbitrary primer. The use of different combinations of G, C and A to anchor the first strand polydT primer enables the priming of the majority of polyadenylated mRNAs. Arbitrary primers may hybridize at none, one or more places along the length of the mRNA, allowing 1st strand cDNA synthesis to occur at none, one or more points in the same gene. In both cases, 2nd strand synthesis is carried out with an arbitrary primer. Since these arbitrary primers for the 2nd strand may also hybridize to the 1st strand cDNA in a number of different places, several different 2nd strand products may be obtained from one binding point of the 1st strand primer. Following 2nd strand synthesis, the original set of primers is used to amplify the second strand products, with the result that numerous gene sequences are amplified.

Restriction endonuclease-facilitated analysis of gene expression

Serial Analysis of Gene Expression (SAGE)

A more recent development in the field of differential display is SAGE analysis (Velculescu *et al.* 1995). This method uses a different approach to those discussed so far and is based on two principles. Firstly, in more than 95% of cases, short nucleotide sequences ('tags') of only nine or 10 base pairs provide sufficient information to identify their gene of origin. Secondly, concatenation (linking together in a series) of these tags allows sequencing of multiple cDNAs within a single clone. Figure 9 shows a schematic representation of the SAGE process. In this procedure, double stranded cDNA from the test cells is synthesized with a biotinylated polydT primer. Following digestion with a commonly cutting (4bp recognition sequence) restriction enzyme ('anchoring enzyme'), the 3' ends of the cDNA population are captured with streptavidin beads. The captured population is

split into two and different adaptors ligated to the 5' ends of each group. Incorporated into the adaptors is a recognition sequence for a type IIS restriction enzyme—one which cuts DNA at a defined distance (< 20 bp) from its recognition sequence. Hence, following digestion of each captured cDNA population with the IIS enzyme, the adaptors plus a short piece of the captured cDNA are released. The two populations are then ligated and the products amplified. The amplified products are cleaved with the original anchoring enzyme, religated (concatomers are formed in the process) and cloned. The advantage of this system is that hundreds of gene tags can be identified by sequencing only a few clones. Furthermore, the number of times a given transcript is identified is a quantitative measurement of that gene's abundance in the original population, a feature which facilitates identification of differentially expressed genes in different cell populations.

Some disadvantages of SAGE analysis include the technical difficulty of the method, a large amount of accurate sequencing is required, biased towards abundant mRNAs, has not been validated in the pharmaco/toxicogenomic setting and has only been used to examine well known tissue differences to date.

Gene Expression Fingerprinting (GEF)

A different capture/restriction digest approach for isolating differentially expressed genes has been described by Ivanova and Belyavsky (1995). In this method, RNA is converted to cDNA using biotinylated oligo(dT) primers. The cDNA population is then digested with a specific endonuclease and captured with magnetic streptavidin microbeads to facilitate removal of the unwanted 5' digestion products. The use of restricted 3'-ends alone serves to reduce the complexity of the cDNA fragment pool and helps to ensure that each RNA species is represented by not more than one restriction product. An adaptor is ligated to facilitate subsequent amplification of the captured population. PCR is carried out with one adaptor-specific and one biotinylated polydT primer. The reamplified population is recaptured and the non-biotinylated strands removed by alkaline dissociation. The non-biotinylated strand is then resynthesized using a different adaptor-specific primer in the presence of a radiolabelled dNTP. The labelled immobilized 3' cDNA ends are next sequentially treated with a series of different restriction endonucleases and the products from each digestion analysed by PAGE. The result is a fingerprint composed of a number of ladders (equal to the number of sequential digests used). By comparing test versus control fingerprints, it is possible to identify differentially expressed products which can then be isolated from the gel and cloned. The advantages of this procedure are that it is very robust and reproducible, and the authors estimate that 80–93% of cDNA molecules are involved in the final fingerprint. The disadvantage is that polyacrylamide gels can rarely resolve more than 300–400 bands, which compares poorly to the 1000 or more which are estimated to be produced in an average experiment. The use of 2-D gels such as those described by Uitterlinden *et al.* (1989) and Hatada *et al.* (1991) may help to overcome this problem.

A similar method for displaying restriction endonuclease fragments was later described by Prashar and Weissman (1996). However, instead of sequential digestion of the immobilized 3'-terminal cDNA fragments, these authors simply compared the profiles of the control and treated populations without further manipulation.

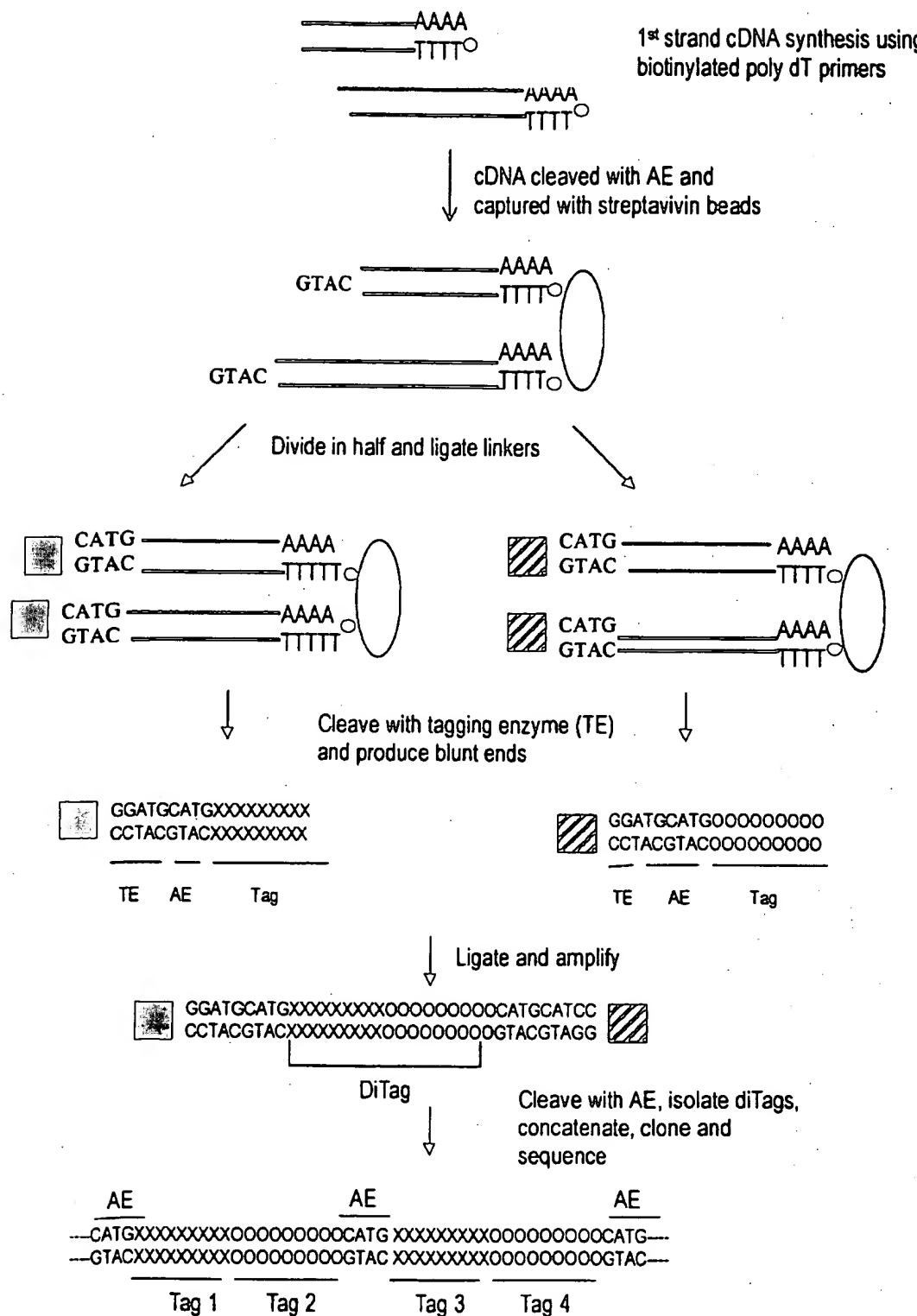


Figure 9. Serial analysis of gene expression (SAGE) analysis. cDNA is cleaved with an anchoring enzyme (AE) and the 3' ends captured using streptavidin beads. The cDNA pool is divided in half and each portion ligated to a different linker, each containing a type IIS restriction site (tagging enzyme, TE). Restriction with the type IIS enzyme releases the linker plus a short length of cDNA (XXXXX and OOOOO indicate nucleotides of different tags). The two pools of tags are then ligated and amplified using linker-specific primers. Following PCR, the products are cleaved with the AE and the diTags isolated from the linkers using PAGE. The diTags are then ligated (during which process, concatenation occurs) and cloned into a vector of choice for sequencing. After Velculescu *et al.* (1995), with permission.

DNA arrays

'Open' differential display systems are cumbersome in that it takes a great deal of time to extract and identify candidate genes and then confirm that they are indeed up- or down-regulated in the treated compared to the control tissue. Normally, the latter process is carried out using Northern blotting or RT-PCR. Even so, each of the aforementioned steps produce a bottleneck to the ultimate goal of rapid analysis of gene expression. These problems will likely be addressed by the development of so-called DNA arrays (e.g. Gress *et al.* 1992, Zhao *et al.* 1995, Schena *et al.* 1996), the introduction of which has signalled the next era in differential gene expression analysis. DNA arrays consist of a gridded membrane or glass 'chips' containing hundreds or thousands of DNA spots, each consisting of multiple copies of part of a known gene. The genes are often selected based on previously proven involvement in oncogenesis, cell cycling, DNA repair, development and other cellular processes. They are usually chosen to be as specific as possible for each gene and animal species. Human and mouse arrays are already commercially available and a few companies will construct a personalized array to order, for example Clontech Laboratories and Research Genetics Inc. The technique is rapid in that hundreds or even thousands of genes can be spotted on a single array, and that mRNA/cDNA from the test populations can be labelled and used directly as probe. When analysed with appropriate hardware and software, arrays offer a rapid and quantitative means to assess differences in gene expression between two cell populations. Of course, there can only be identification and quantitation of those genes which are in the array (hence the term 'closed' system). Therefore, one approach to elucidating the molecular mechanisms involved in a particular disease/development system may be to combine an open and closed system—a DNA array to directly identify and quantitate the expression of known genes in mRNA populations, and an open system such as SSH to isolate unknown genes which are differentially expressed.

One of the main advantages of DNA arrays is the huge number of gene fragments which can be put on a membrane—some companies have reported gridding up to 60 000 spots on a single glass 'chip' (microscope slide). These high density chip-based micro-arrays will probably become available as mass-produced off-the-shelf items in the near future. This should facilitate the more rapid determination of differential expression in time and dose-response experiments. Aside from their high cost and the technical complexities involved in producing and probing DNA arrays, the main problem which remains, especially with the newer micro-array (gene-chip) technologies, is that results are often not wholly reproducible between arrays. However, this problem is being addressed and should be resolved within the next few years.

EST databases as a means to identify differentially expressed genes

Expressed sequence tags (ESTs) are partial sequences of clones obtained from cDNA libraries. Even though most ESTs have no formal identity (putative identification is the best to be hoped for), they have proven to be a rapid and efficient means of discovering new genes and can be used to generate profiles of gene-expression in specific cells. Since they were first described by Adams *et al.* (1991), there has been a huge explosion in EST production and it is estimated that there are now well over a million such sequences in the public domain, representing over half

of all human genes (Hillier *et al.* 1996). This large number of freely available sequences (both sequence information and clones are normally available royalty-free from the originators) has enabled the development of a new approach towards differential gene expression analysis as described by Vasmatazis *et al.* (1998). The approach is simple in theory: EST databases are first searched for genes that have a number of related EST sequences from the target tissue of choice, but none or few from non-target tissue libraries. Programmes to assist in the assembly of such sets of overlapping data may be developed in-house or obtained privately or from the internet. For example, the Institute for Genomic Research (TIGR, found at <http://www.tigr.org>) provides many software tools free of charge to the scientific community. Included amongst these is the TIGR assembler (Sutton *et al.* 1995), a tool for the assembly of large sets of overlapping data such as ESTs, bacterial artificial chromosomes (BAC)s, or small genomes. Candidate EST clones representing different genes are then analysed using RNA blot methods for size and tissue specificity and, if required, used as probes to isolate and identify the full length cDNA clone for further characterization. In practice however, the method is rather more involved, requiring bioinformatic and computer analysis coupled with confirmatory molecular studies. Vasmatazis *et al.* (1998) have described several problems in this fledgling approach, such as separating highly homologous sequences derived from different genes and an overemphasis of specificity for some EST sequences. However, since these problems will largely be addressed by the development of more suitable computer algorithms and an increased completeness of the EST database, it is likely that this approach to identifying differentially expressed genes may enjoy more patronage in the future.

Problems and potential of differential expression techniques

The holistic or single cell approach?

When working with *in vivo* models of differential expression, one of the first issues to consider must be the presence of multiple cell types in any given specimen. For example, a liver sample is likely to contain not only hepatocytes, but also (potentially) Ito cells, bile ductule cells, endothelial cells, various immune cells (e.g. lymphocytes, macrophages and Kupffer cells) and fibroblasts. Other tissues will each have their own distinctive cell populations. Also, in the case of neoplastic tissue, there are almost always normal, hyperplastic and/or dysplastic cells present in a sample. One must, therefore, be aware that genes obtained from a differential display experiment performed on an animal tissue model may not necessarily arise exclusively from the intended 'target' cells, e.g. hepatocytes/neoplastic cells. If appropriate, further analyses using immunohistochemistry, *in situ* hybridization or *in situ* RT-PCR should be used to confirm which cell types are expressing the gene(s) of interest. This problem is probably most acute for those studying the differential expression of genes in the development of different cell types, where there is a need to examine homologous cell populations. The problem is now being addressed at the National Cancer Institute (Bethesda, MD, USA) where new microdissection techniques have been employed to assist in their gene analysis programme, the Cancer Genome Anatomy Project (CGAP) (For more information see web site: <http://www.ncbi.nlm.nih.gov/ncicgap/intro.html>). There are also separation techniques available that utilise cell-specific antigens as a means to isolate target cells,

e.g. fluorescence activated cell sorting (FACS) (Dunbar *et al.* 1998, Kas-Deelen *et al.* 1998) and magnetic bead technology (Richard *et al.* 1998, Rogler *et al.* 1998).

However, those taking a holistic approach may consider this issue unimportant. There is an equally appropriate view that all those genes showing altered expression within a compromised tissue should be taken into consideration. After all, since all tissues are complex mixes of different, interacting cell types which intimately regulate each other's growth and development, it is clear that each cell type could in some way contribute (positively or negatively) towards the molecular mechanisms which lie behind responses to external stimuli or neoplastic growth. It is perhaps then more informative to carry out differential display experiments using *in vivo* as opposed to *in vitro* models, where uniform populations of identical cells probably represent a partial, skewed or even inaccurate picture of the molecular changes that occur.

The incidence and possible implications of inter-individual biological variation should be considered in any approach where whole animal models are being used. It is clear that individuals (humans and animals) respond in different ways to identical stimuli. One of the best characterized examples is the debrisoquine oxidation polymorphism, which is mediated by cytochrome CYP2D6 and determines the pharmacokinetics of many commonly prescribed drugs (Lennard 1993, Meyer and Zanger 1997). The reasons for such differences are varied and complex, but allelic variations, regulatory region polymorphisms and even physical and mental health can all contribute to observed differences in individual responses. Careful thought should, therefore, be given to the specific objectives of the study and to the possible value of pooling starting material (tissue/mRNA). The effect of this can be beneficial through the ironing out of exaggerated responses and unimportant minor fluctuations of (mechanistically) irrelevant genes in individual animals, thus providing a clearer overall picture of the general molecular mechanisms of the response. However, at the same time such minor variations may be of utmost importance in deciding the ability of individual animals to succumb to or resist the effects of a given chemical/disease.

How efficient are differential expression techniques at recovering a high percentage of differentially expressed genes?

A number of groups have produced experimental data suggesting that mammalian cells produce between 8000–15 000 different mRNA species at any one time (Mechler and Rabbitts 1981, Hedrick *et al.* 1984, Bravo 1990), although figures as high as 20–30 000 have also been quoted (Axel *et al.* 1976). Hedrick *et al.* (1984) provided evidence suggesting that the majority of these belong to the rare abundance class. A breakdown of this abundance distribution is shown in table 1.

When the results of differential display experiments have been compared with data obtained previously using other methods, it is apparent that not all differentially expressed mRNAs are represented in the final display. In particular, rare messages (which, importantly, often include regulatory proteins) are not easily recovered using differential display systems. This is a major shortcoming, as the majority of mRNA species exist at levels of less than 0.005% of the total population (table 1). Bertoli *et al.* (1995) examined the efficiency of DD templates (heterogeneous mRNA populations) for recovering rare messages and were unable to detect mRNA

species present at less than 1.2% of the total mRNA population—equivalent to an intermediate or abundant species. Interestingly, when simple model systems (single target only) were used instead of a heterogeneous mRNA population, the same primers could detect levels of target mRNA down to 10 000× smaller. These results are probably best explained by competition for substrates from the many PCR products produced in a DD reaction.

The numbers of differentially expressed mRNAs reported in the literature using various model systems provides further evidence that many differentially expressed mRNAs are not recovered. For example, DeRisi *et al.* (1997) used DNA array technology to examine gene expression in yeast following exhaustion of sugar in the medium, and found that more than 1700 genes showed a change in expression of at least 2-fold. In light of such a finding, it would not be unreasonable to suggest that of the 8000–15 000 different mRNA species produced by any given mammalian cell, up to 1000 or more may show altered expression following chemical stimulation. Whilst this may be an extreme figure, it is known that at least 100 genes are activated/upregulated in Jurkat (T-) cells following IL-2 stimulation (Ullman *et al.* 1990). In addition, Wan *et al.* (1996) estimated that interferon- γ -stimulated HeLa cells differentially express up to 433 genes (assuming 24 000 distinct mRNAs expressed by the cells). However, there have been few publications documenting anywhere near the recovery of these numbers. For example, in using DD to compare normal and regenerating mouse liver, Bauer *et al.* (1993) found only 70 of 38 000 total bands to be different. Of these, 50% (35 genes) were shown to correspond to differentially expressed bands. Chen *et al.* (1996) reported 10 genes upregulated in female rat liver following ethinyl estradiol treatment. McKenzie and Drake (1997) identified 14 different gene products whose expression was altered by phorbol myristate acetate (PMA, a tumour promoter agent) stimulation of a human myelomonocytic cell line. Kilty and Vickers (1997) identified 10 different gene products whose expression was upregulated in the peripheral blood leukocytes of allergic disease sufferers. Linskens *et al.* (1995) found 23 genes differentially expressed between young and senescent fibroblasts. Techniques other than DD have also provided an apparent paucity of differentially expressed genes. Using SH for example, Cao *et al.* (1997) found 15 genes differentially expressed in colorectal cancer compared to normal mucosal epithelium. Fitzpatrick *et al.* (1995) isolated 17 genes upregulated in rat liver following treatment with the peroxisome proliferator, clofibrate; Philips *et al.* (1990) isolated 12 cDNA clones which were upregulated in highly metastatic mammary adenocarcinoma cell lines compared to poorly metastatic ones. Prashar and Weissman (1996) used 3' restriction fragment analysis and identified approximately 40 genes showing altered expression within 4 h of activation of Jurkat T-cells. Groenink and Leegwater (1996) analysed 27 gene fragments isolated using SSH of delayed early response phase of liver regeneration and found only 12 to be upregulated.

In the laboratory, SSH was used to isolate up to 70 candidate genes which appear to show altered expression in guinea pig liver following short-term treatment with the peroxisome proliferator, WY-14,643 (Rockett, Swales, Esdaile and Gibson, unpublished observations). However, these findings have still to be confirmed by analysis of the extracted tissue mRNA for differential expression of these sequences.

Whilst the latest differential display technologies are purported to include design and experimental modifications to overcome this lack of efficiency (in both the total number of differentially expressed genes recovered and the percentage that are true

positives), it is still not clear if such adaptations are practically effective—proving efficiency by spiking with a known amount of limited numbers of artificial construct(s) is one thing, but isolating a high percentage of the rare messages already present in an mRNA population is another. Of course, some models will genuinely produce only a small number of differentially expressed genes. In addition, there are also technical problems that can reduce efficiency. For example, mRNAs may have an unusual primary structure that effectively prevents their amplification by PCR-based systems. In addition, it is known that under certain circumstances not all mRNAs have 3' polyA sites. For example, during *Xenopus* development, deadenylation is used as a means to stabilize RNAs (Voeltz and Steitz 1998), whilst preferential deadenylation may play a role in regulating Hsp70 (and perhaps, therefore, other stress protein) expression in *Drosophila* (Dellavalle *et al.* 1994). The presence of deadenylated mRNAs would clearly reduce the efficiency of systems utilizing a polydT reverse transcription step. The efficiency of any system also depends on the quality of the starting material. All differential display techniques use mRNA as their target material. However, it is difficult to isolate mRNA that is completely free of ribosomal RNA. Even if polydT primers are used to prime first strand cDNA synthesis, ribosomal RNA is often transcribed to some degree (Clontech PCR-Select cDNA Subtraction kit user manual). It has been shown, at least in the case of SSH, that a high rRNA:mRNA ratio can lead to inefficient subtractive hybridization (Clontech PCR-Select cDNA Subtraction kit user manual), and there is no reason to suppose that it will not do likewise in other SH approaches. Finally, those techniques that utilise a presubtraction amplification step (e.g. RDA) may present a skewed representation since some sequences amplify better than others.

Of course, probably the most important consideration is the temporal factor. It is clear that any given differential display experiment can only interrogate a cell at one point in time. It may well be that a high percentage of the genes showing altered expression at that time are obtained. However, given that disease processes and responses to environmental stimuli involve dynamic cascades of signalling, regulation, production and action, it is clear that all those genes which are switched on/off at different times will not be recovered and, therefore, vital information may well be missed. It is, therefore, imperative to obtain as much information about the model system beforehand as possible, from which a strategy can be derived for targeting specific time points or events that are of particular interest to the investigator. One way of getting round this problem of single time point analysis is to conduct the experiment over a suitable time course which, of course, adds substantially to the amount of work involved.

How sensitive are differential expression technologies?

There has been little published data that addresses the issue of how large the change in expression must be for it to permit isolation of the gene in question with the various differential expression technologies. Although the isolation of genes whose expression is changed as little as 1.5-fold has been reported using SSH (Groenink and Leegwater 1996), it appears that those demonstrating a change in excess of 5-fold are more likely to be picked up. Thus, there is a 'grey zone' in between where small changes could fade in and out of isolation between

experiments and animals. DD, on the other hand, is not subject to this grey zone since, unlike SH approaches, it does not amplify the difference in expression between two samples. Wan *et al.* (1996) reported that differences in expression of twofold or more are detectable using DD.

Resolution and visualization of differential expression products

It seems highly improbable with current technology that a gel system could be developed that is able to resolve all gene species showing altered expression in any given test system (be it SH- or DD-based). Polyacrylamide gel electrophoresis (PAGE) can resolve size differences down to 0.2% (Sambrook *et al.* 1989) and are used as standard in DD experiments. Even so, it is clear that a complex series of gene products such as those seen in a DD will contain unresolvable components. Thus, what appears to be one band in a gel may in fact turn out to be several. Indeed, it has been well documented (Mathieu-Daude *et al.* 1996, Smith *et al.* 1997) that a single band extracted from a DD often represents a composite of heterogeneous products, and the same has been found for SSH displays in this laboratory (Rockett *et al.* 1997). One possible solution was offered by Mathieu-Daude *et al.* (1996), who extracted and reamplified candidate bands from a DD display and used single strand conformation polymorphism (SSCP) analysis to confirm which components represented the truly differentially expressed product.

Many scientists often try to avoid the use of PAGE where possible because it is technically more demanding than agarose gel electrophoresis (AGE). Unfortunately, high resolution agarose gels such as Metaphor (FMC, Lichfield, UK) and AquaPor HR (National Diagnostics, Hesse, UK), whilst easier to prepare and manipulate than PAGE, can only separate DNA sequences which differ in size by around 1.5–2% (15–20 base pairs for a 1Kb fragment). Thus, SSH, RDA or other such products which differ in size by less than this amount are normally not resolvable. However, a simple technique does in fact exist for increasing the resolving power of AGE—the inclusion of HA-red (10-phenyl neutral red-PEG ligand) or HA-yellow (bisbenzamide-PEG ligand) (Hanse Analytik GmbH, Bremen, Germany) in a gel separates identical or closely sized products on base content. Specifically, HA-red and -yellow selectively bind to GC and AT DNA motifs, respectively (Wawer *et al.* 1995, Hanse Analytik 1997, personal communication). Since both HA-stains possess an overall positive charge, they migrate towards the cathode when an electric field is applied. This is in direct opposition to DNA, which is negatively charged and, therefore, migrates towards the anode. Thus, if two DNA clones are identical in size (as perceived on a standard high resolution agarose gel), but differ in AT/GC content, inclusion of a HA-dye in the gel will effectively retard the migration of one of the sequences compared to the other, effectively making it apparently larger and, thus, providing a means of differentiating between the two. The use of HA-red has been shown to resolve sequences with an AT variation of less than 1% (Wawer *et al.* 1995), whilst Hanse Analytik have reported that HA staining is so sensitive that in one case it was used to distinguish two 567bp sequences which differed by only a single point mutation (Hanse Analytik 1996, personal communication). Therefore, if one wishes to check whether all the clones produced from a specific band in a differential display experiment are derived from the same gene species, a small amount of reamplified or digested clone can be run on a standard high resolution gel, and a second aliquot

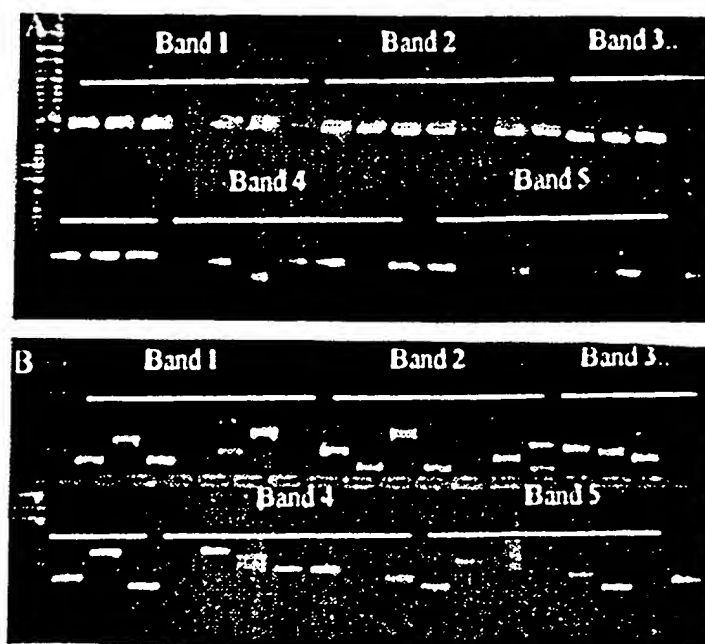


Figure 10. Discrimination of clones of identical/nearly identical size using HA-red. Bands of decreasing size (1–5) were extracted from the final display of a suppression subtractive hybridization experiment and cloned. Seven colonies were picked at random from each cloned band and their inserts amplified using PCR. The products were run on two gels, (A) a high resolution 2% agarose gel, and (B) a high resolution 2% agarose gel containing 1 U/ml HA-red. With few exceptions, all the clones from each band appear to be the same size (gel A). However, the presence of HA-red (gel B), which separates identically-sized DNA fragments based on the percentage of GC within the sequence, clearly indicates the presence of different gene species within each band. For example, even though all five re-amplified clones of band 1 appear to be the same size, at least four different gene species are represented.

in a similar gel containing one of the HA-stains. The standard gel should indicate any gross size differences, whilst the HA-stained gel should separate otherwise unresolvable species (on standard AGE) according to their base content. Geisinger *et al.* (1997) reported successful use of this approach for identifying DD-derived clones. Figure 10 shows such an experiment carried out in this laboratory on clones obtained from a band extracted from an SSH display.

An alternative approach is to carry out a 2-D analysis of the differential display products. In this approach, size-based separation is first carried out in a standard agarose gel. The gel slice containing the display is then extracted and incorporated in to a HA gel for resolution based on AT/GC content.

Of course, one should always consider the possibility of there being different gene species which are the same size and have the same GC/AT content. However, even these species are not unresolvable given some effort—again, one might use SSCP, or perhaps a denaturing gradient gel electrophoresis (DGGE) or temperature gradient field electrophoresis (TGGE) approach to resolve the contents of a band, either directly on the extracted band (Suzuki *et al.* 1991) or on the reamplified product.

The requirement of some differential display techniques to visualize large numbers of products (e.g. DD and GEF) can also present a problem in that, in terms of numbers, the resolution of PAGE rarely exceeds 300–400 bands. One approach to overcoming this might be to use 2-D gels such as those described by Uitterlinden *et al.* (1989) and Hatada *et al.* (1991).

Extraction of differentially expressed bands from a gel can be complex since, in some cases (e.g. DD, GEF), the results are visualized by autoradiographic means, such that precise overlay of the developed film on the gel must occur if the correct band is to be extracted for further analysis. Clearly, a misjudged extraction can account for many man-hours lost. This problem, and that of the use of radioisotopes, has been addressed by several groups. For example, Lohmann *et al.* (1995) demonstrated that silver staining can be used directly to visualize DD bands in horizontal PAGs. An *et al.* (1996) avoided the use of radioisotopes by transferring a small amount (20–30%) of the DNA from their DD to a nylon membrane, and visualizing the bands using chemiluminescent staining before going back to extract the remaining DNA from the gel. Chen and Peck (1996) went one step further and transferred the entire DD to a nylon membrane. The DNA bands were then visualized using a digoxigenin (DIG) system (DIG was attached to the polydT primers used in the differential display procedure). Differentially expressed bands were cut from the membrane and the DNA eluted by washing with PCR buffer prior to reamplification.

One of the advantages of using techniques such as SSH and RDA is that the final display can be run on an agarose gel and the bands visualized with simple ethidium bromide staining. Whilst this approach can provide acceptable results, overstaining with SYBR Green I or SYBR Gold nucleic acid stains (FMC) effectively enhances the intensity and sharpness of the bands. This greatly aids in their precise extraction and often reveals some faint products that may otherwise be overlooked. Whilst differential displays stained with SYBR Green I are better visualized using short wavelength UV (254 nm) rather than medium wavelength (306 nm), the shorter wavelength is much more DNA damaging. In practice, it takes only a few seconds to damage DNA extracted under 254 nm irradiation, effectively preventing reamplification and cloning. The best approach is to over stain with SYBR Green I and extract bands under a medium wavelength UV transillumination.

The possible use of 'microfingerprinting' to reduce complexity

Given the sheer number of gene products and the possible complexity of each band, an alternative approach to rapid characterization may be to use an enhanced analysis of a small section of a differential display—a 'sub-fingerprint' or 'micro-fingerprint'. In this case, one could concentrate on those bands which only appear in a particular chosen size region. Reducing the fingerprint in this way has at least two advantages. One is that it should be possible to use different gel types, concentrations and run times tailored exactly to that region. Currently, one might run products from 100–3000 + bp on the same gel, which leads to compromise in the gel system being used and consequently to suboptimal resolution, both in terms of size and numbers, and can lead to problems in the accurate excision of individual bands. Secondly, it may be possible to enhance resolution by using a 2-D analysis using a HA-stain, as described earlier. In summary, if a range of gene product sizes is carefully chosen to include certain 'relevant' genes, the 2-D system standardized, and appropriate gene analysis used, it may be possible to develop a method for the early and rapid identification of compounds which have similar or widely different cellular effects. If the prognosis for exposure to one or more other chemicals which display a similar profile is already known, then one could perhaps predict similar effects for any new compounds which show a similar micro-fingerprint.

An alternative approach to microfingerprinting is to examine altered expression in specific families of genes through careful selection of PCR primers and/or post-reaction analysis. Stress genes, growth factors and/or their receptors, cell cycling genes, cytochromes P450 and regulatory proteins might be considered as candidates for analysis in this way. Indeed, some off-the-shelf DNA arrays (e.g. Clontech's Atlas cDNA Expression Array series) already anticipated this to some degree by grouping together genes involved in different responses e.g. apoptosis, stress, DNA-damage response etc.

Screening

False positives

The generation of false positives has been discussed at length amongst the differential display community (Liang *et al.* 1993, 1995, Nishio *et al.* 1994, Sun *et al.* 1994, Sompayrac *et al.* 1995). The reason for false positives varies with the technique being used. For instance, in RDA, the use of adaptors which have not been HPLC purified can lead to the production of false positives through illegitimate ligation events (O'Neill and Sinclair 1997), whilst in DD they can arise through PCR artifacts and illegitimate transcription of rRNA. In SH, false positives appear to be derived largely from abundant gene species, although some may arise from cDNA/mRNA species which do not undergo hybridization for technical reasons.

A quick screening of putative differentially expressed clones can be carried out using a simple dot blot approach, in which labelled first strand probes synthesized from tester and driver mRNA are hybridized to an array of said clones (Hedrick *et al.* 1984, Sakaguchi *et al.* 1986). Differentially expressed clones will hybridize to tester probe, but not driver. The disadvantage of this approach is that rare species may not generate detectable hybridization signals. One option for those using SSH is to screen the clones using a labelled probe generated from the subtracted cDNA from which it was derived, and with a probe made from the reverse subtraction reaction (ClonTechniques 1997a). Since the SSH method enriches rare sequences, it should be possible to confirm the presence of clones representing low abundance genes. Despite this quick screening step, there is still the need to go back to the original mRNA and confirm the altered expression using a more quantitative approach. Although this may be achieved using Northern blots, the sensitivity is poor by today's high standards and one must rely on PCR methods for accurate and sensitive determinations (see below).

Sequence analysis

The majority of differential display procedures produce final products which are between 100 and 1000bp in size. However, this may considerably reduce the size of the sequence for analysis of the DNA databases. This in turn leads to a reduced confidence in the result—several families of genes have members whose DNA sequences are almost identical except in a few key stretches, e.g. the cytochrome P450 gene superfamily (Nelson *et al.* 1996). Thus, does the clone identified as being almost identical to gene X_0 really come from that gene, or its brother gene X_1 , or its as yet undiscovered sister X_2 ? For example, using SSH, part of a gene was isolated,

which was up-regulated in the liver of rats exposed to Wy-14,643 and was identified by a FASTA search as being transferrin (data not shown). However, transferrin is known to be downregulated by hypolipidemic peroxisome proliferators such as Wy-14,643 (Hertz *et al.* 1996), and this was confirmed with subsequent RT-PCR analysis. This suggests that the gene sequence isolated may belong to a gene which is closely related to transferrin, but is regulated by a different mechanism.

A further problem associated with SH technology is redundancy. In most cases before SH is carried out, the cDNA population must first be simplified by restriction digestion. This is important for at least two reasons:

- (1) To reduce complexity—long cDNA fragments may form complex networks which prevent the formation of appropriate hybrids, especially at the high concentrations required for efficient hybridization.
- (2) Cutting the cDNAs into small fragments provides better representation of individual genes. This is because genes derived from related but distinct members of gene families often have similar coding sequences that may cross-hybridize and be eliminated during the subtraction procedure (Ko 1990). Furthermore, different fragments from the same cDNA may differ considerably in terms of hybridization and amplification and, thus, may not efficiently do one or the other (Wang and Brown 1991). Thus, some fragments from differentially expressed cDNAs may be eliminated during subtractive hybridization procedures. However, other fragments may be enriched and isolated. As a consequence of this, some genes will be cut one or more times, giving rise to two or more fragments of different sizes. If those same genes are differentially expressed, then two or more of the different size fragments may come through as separate bands on the final differential display, increasing the observed redundancy and increasing the number of redundant sequencing reactions.

Sequence comparisons also throw up another important point—at what degree of sequence similarity does one accept a result. Is 90% identity between a gene derived from your model species and another acceptably close? Is 95% between your sequence and one from the same species also acceptable? This problem is particularly relevant when the forward and reverse sequence comparisons give similar sequences with completely different gene species! An arbitrary decision seems to be to allocate genes that are definite (95% and above similarity) and then group those between 60 and 95% as being related or possible homologues.

Quantitative analysis

At some point, one must give consideration to the quantitative analysis of the candidate genes, either as a means of confirming that they are truly differentially expressed, or in order to establish just what the differences are. Northern blot analysis is a popular approach as it is relatively easy and quick to perform. However, the major drawback with Northern blots is that they are often not sensitive enough to detect rare sequences. Since the majority of messages expressed in a cell are of low abundance (see table 1), this is a major problem. Consequently, RT-PCR may be the method of choice for confirming differential expression. Although the procedure is somewhat more complex than Northern analysis, requiring synthesis of primers and optimization of reaction conditions for each gene species, it is now possible to set up high throughput PCR systems using multichannel pipettes, 96 +well plates and

appropriate thermal cycling technology. Whilst quantitative analysis is more desirable, being more accurate and without reliance on an internal standard, the money and time needed to develop a competitor molecule is often excessive, especially when one might be examining tens or even hundreds of gene species. The use of semi-quantitative analysis is simpler, although still relatively involved. One must first of all choose an internal standard that does not change in the test cells compared to the controls. Numerous reference genes have been tried in the past, for example interferon-gamma (IFN- γ , Frye *et al.* 1989), β -actin (Heuval *et al.* 1994), glyceraldehyde-3-phosphate dehydrogenase (GAPDH, Wong *et al.* 1994), dihydrofolate reductase (DHFR, Mohler and Butler 1991), β -2-microglobulin (β -2-m, Murphy *et al.* 1990), hypoxanthine phosphoribosyl transferase (HPRT, Foss *et al.* 1998) and a number of others (ClonTechniques 1997b). Ideally, an internal standard should not change its level of expression in the cell regardless of cell age, stage in the cell cycle or through the effects of external stimuli. However, it has been shown on numerous occasions that the levels of most housekeeping genes currently used by the research community do in fact change under certain conditions and in different tissues (ClonTechniques 1997b). It is imperative, therefore, that preliminary experiments be carried out on a panel of housekeeping genes to establish their suitability for use in the model system.

Interpretation of quantitative data must also be treated with caution. By comparing the lists of genes identified by differential expression one can perhaps gain insight into why two different species react in different ways to external stimuli. For example, rats and mice appear sensitive to the non-genotoxic effects of a wide range of peroxisome proliferators whilst Syrian hamsters and guinea pigs are largely resistant (Orton *et al.* 1984, Rodricks and Turnbull 1987, Lake *et al.* 1989, 1993, Makowska *et al.* 1992). A simplified approach to resolving the reason(s) why is to compare lists of up- and down-regulated genes in order to identify those which are expressed in only one species and, through background knowledge of the effects of the said gene, might suggest a mechanism of facilitated non-genotoxic carcinogenesis or protection. Of course, the situation is likely to be far more complex. Perhaps if there were one key gene protecting guinea pig from non-genotoxic effects and it was upregulated 50 times by PPs, the same gene might only be up-regulated five times in the rat. However, since both were noted to be upregulated, the importance of the gene may be overlooked. Just to complicate matters, a large change in expression does not necessarily mean a biologically important change. For example, what is the true relevance of gene Y which shows a 50-fold increase after a particular treatment, and gene Z which shows only a 5-fold increase? If one examines the literature one may find that historically, gene Y has often been shown to be up-regulated 40–60-fold by a number of unrelated stimuli—in light of this the 50-fold increase would appear less significant. However, the literature may show that gene Z has never been recorded as having more than doubled in expression—which makes your 5-fold increase all the more exciting. Perhaps even more interesting is if that same 5-fold increase has only been seen in related neoplasms or following treatment with related chemicals.

Problems in using the differential display approach

Differential display technology originally held promise of an easily obtainable 'fingerprint' of those genes which are up- or down-regulated in test animals/cells in a developmental process or following exposure to given stimuli. However, it has

become clear that the fingerprinting process, whilst still valid, is much too complex to be represented by a single technique profile. This is because all differential display techniques have common and/or unique technical problems which preclude the isolation and identification of all those genes which show changes in expression. Furthermore, there are important genetic changes related to disease development which differential expression analysis is simply not designed to address. An example of this is the presence of small deletions, insertions, or point mutations such as those seen in activated oncogenes, tumour suppressor genes and individual polymorphisms. Polymorphic variations, small though they usually are, are often regarded as being of paramount importance in explaining why some patients respond better than others to certain drug treatments (and, in logical extension, why some people are less affected by potentially dangerous xenobiotics/carcinogens than others). The identification of such point mutations and naturally occurring polymorphisms requires the subsequent application of sequencing, SSCP, DGGE or TGGE to the gene of interest. Furthermore, differential display is not designed to address issues such as alternatively spliced gene species or whether an increased abundance of mRNA is a result of increased transcription or increased mRNA stability.

Conclusions

Perhaps the main advantage of open system differential display techniques is that they are not limited by extant theories or researcher bias in revealing genes which are differentially expressed, since they are designed to amplify all genes which demonstrate altered expression. This means that they are useful for the isolation of previously unknown genes which may turn out be useful biomarkers of a particular state or condition. At least one open system (SAGE) is also quantitative, thus eliminating the need to return to the original mRNA and carry out Northern/PCR analysis to confirm the result. However, the rapid progress of genome mapping projects means that over the next 5–10 years or so, the balance of experimental use will switch from open to closed differential display systems, particularly DNA arrays. Arrays are easier and faster to prepare and use, provide quantitative data, are suitable for high throughput analysis and can be tailored to look at specific signalling pathways or families of genes. Identification of all the gene sequences in human and common laboratory animals combined with improved DNA array technology, means that it will soon no longer be necessary to try to isolate differentially expressed genes using the technically more demanding open system approach. Thus, their main advantage (that of identifying unknown genes) will be largely eradicated. It is likely, therefore, that their sphere of application will be reduced to analysis of the less common laboratory species, since it will be some time yet before the genomes of such animals as zebrafish, electric eels, gerbils, crayfish and squid, for example, will be sequenced.

Of course, in the end the question will always remain: What is the functional/biological significance of the identified, differentially expressed genes? One persistent problem is understanding whether differentially expressed genes are a cause or consequence of the altered state. Furthermore, many chemicals, such as non-genotoxic carcinogens, are also mitogens and so genes associated with replication will also be upregulated but may have little or nothing to do with the

carcinogenic effect. Whilst differential display technology cannot hope to answer these questions, it does provide a springboard from which identification, regulatory and functional studies can be launched. Understanding the molecular mechanism of cellular responses is almost impossible without knowing the regulation and function of those genes and their condition (e.g. mutated). In an abstract sense, differential display can be likened to a still photograph, showing details of a fixed moment in time. Consider the Historian who knows the outcome of a battle and the placement and condition of the troops before the battle commenced, but is asked to try and deduce how the battle progressed and why it ended as it did from a few still photographs—an impossible task. In order to understand the battle, the Historian must find out the capabilities and motivation of the soldiers and their commanding officers, what the orders were and whether they were obeyed. He must examine the terrain, the remains of the battle and consider the effects the prevailing weather conditions exerted. Likewise, if mechanistic answers are to be forthcoming, the scientist must use differential display in combination with other techniques, such as knockout technology, the analysis of cell signalling pathways, mutation analysis and time and dose response analyses. Although this review has emphasized the importance of differential gene profiling, it should not be considered in isolation and the full impact of this approach will be strengthened if used in combination with functional genomics and proteomics (2-dimensional protein gels from isoelectric focusing and subsequent SDS electrophoresis and virtual 2D-maps using capillary electrophoresis). Proteomics is attracting much recent attention as many of the changes resulting in differential gene expression do not involve changes in mRNA levels, as described extensively herein, but rather protein-protein, protein-DNA and protein phosphorylation events which would require functional genomics or proteomic technologies for investigation.

Despite the limitations of differential display technology, it is clear that many potential applications and benefits can be obtained from characterizing the genetic changes that occur in a cell during normal and disease development and in response to chemical or biological insult. In light of functional data, such profiling will provide a 'fingerprint' of each stage of development or response, and in the long term should help in the elucidation of specific and sensitive biomarkers for different types of chemical/biological exposure and disease states. The potential medical and therapeutic benefits of understanding such molecular changes are almost immeasurable. Amongst other things, such fingerprints could indicate the family or even specific type of chemical an individual has been exposed to plus the length and/or acuteness of that exposure, thus indicating the most prudent treatment. They may also help uncover differences in histologically identical cancers, provide diagnostic tests for the earliest stages of neoplasia and, again, perhaps indicate the most efficacious treatment.

The Human Genome Project will be completed early in the next century and the DNA sequence of all the human genes will be known. The continuing development and evolution of differential gene expression technology will ensure that this knowledge contributes fully to the understanding of human disease processes.

Acknowledgements

We acknowledge Drs Nick Plant (University of Surrey), Sally Darney and Chris Luft (US EPA at RTP) for their critical analysis of the manuscript prior to submission. This manuscript has been reviewed in accordance with the policy of the

US Environmental Protection Agency and approved for publication. Approval does not signify that the contents reflect the views and policies of the Agency, nor does mention of trade names constitute endorsement or recommendation for use.

References

- ADAMS, M. D., KELLEY, J. M., GOCAYNE, J. D., DUBNICK, M., POLYMERPOULOS, M. H., XIAO, H., MERRIL, C. R., WU, A., OLDE, B., MORENO, R. F., KERLAVAGE, A. R., MCCOMBIE, W. R. and VENTOR, J. C., 1991, Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, **252**, 1651-1656.
- AN, G., LUO, G., VELTRI, R. W. and O'HARA, S. M., 1996, Sensitive non-radioactive differential display method using chemiluminescent detection. *Biotechniques*, **20**, 342-346.
- AXEL, R., FEIGELSON, P. and SCHULTZ, G., 1976, Analysis of the complexity and diversity of mRNA from chicken liver and oviduct. *Cell*, **7**, 247-254.
- BAND, V. and SAGER, R., 1989, Distinctive traits of normal and tumor-derived human mammary epithelial cells expressed in a medium that supports long-term growth of both cell types. *Proceedings of the National Academy of Sciences, USA*, **86**, 1249-1253.
- BAUER, D., MULLER, H., REICH, J., RIEDEL, H., AHRENKIEL, V., WARTHOF, P. and STRAUSS, M., 1993, Identification of differentially expressed mRNA species by an improved display technique (DDRT-PCR). *Nucleic Acids Research*, **21**, 4272-4280.
- BERTIOLI, D. J., SCHLICHTER, U. H. A., ADAMS, M. J., BURROWS, P. R., STEINBISS, H.-H. and ANTONIW, J. F., 1995, An analysis of differential display shows a strong bias towards high copy number mRNAs. *Nucleic Acids Research*, **23**, 4520-4523.
- BRAVO, R., 1990, Genes induced during the G0/G1 transition in mouse fibroblasts. *Seminars in Cancer Biology*, **1**, 37-46.
- BURN, T. C., PETROVICK, M. S., HOHAUS, S., ROLLINS, B. J. and TENEN, D. G., 1994, Monocyte chemoattractant protein-1 gene is expressed in activated neutrophils and retinoic acid-induced human myeloid cell lines. *Blood*, **84**, 2776-2783.
- CAO, J., CAI, X., ZHENG, L., GENG, L., SHI, Z., PAO, C. C. and ZHENG, S., 1997, Characterisation of colorectal cancer-related cDNA clones obtained by subtractive hybridisation screening. *Journal of Cancer Research and Clinical Oncology*, **123**, 447-451.
- CASSIDY, S. B., 1995, Uniparental disomy and genomic imprinting as causes of human genetic disease. *Environmental and Molecular Mutagenesis*, **25** (Suppl 26), 13-20.
- CHANG, G. W. and TERZAGHI-HOWE, M., 1998, Multiple changes in gene expression are associated with normal cell-induced modulation of the neoplastic phenotype. *Cancer Research*, **58**, 4445-4452.
- CHEN, J., SCHWARTZ, D. A., YOUNG, T. A., NORRIS, J. S. and YAGER, J. D., 1996, Identification of genes whose expression is altered during mitosuppression in livers of ethinyl estradiol-treated female rats. *Carcinogenesis*, **17**, 2783-2786.
- CHEN, J. J. W. and PECK, K., 1996, Non-radioactive differential display method to directly visualise and amplify differential bands on nylon membrane. *Nucleic Acid Research*, **24**, 793-794.
- CLON TECHNIQUES, 1997a, PCR-Select Differential Screening Kit—the nextstep after Clontech PCR-Select cDNA subtraction. *ClonTechniques*, **XII**, 18-19.
- CLON TECHNIQUES, 1997b, Housekeeping RT-PCR amplimers and cDNA probes. *ClonTechniques*, **XII**, 15-16.
- DAVIS, M. M., COHEN, D. I., NIELSEN, E. A., STEINMETZ, M., PAUL, W. E. and HOOD, L., 1984, Cell-type-specific cDNA probes and the murine I region: the localization and orientation of Ad alpha. *Proceedings of the National Academy of Sciences (USA)*, **81**, 2194-2198.
- DELLAVALLE, R. P., PETERSON, R. and LINDQUIST, S., 1994, Preferential deadenylation of HSP70 mRNA plays a key role in regulating Hsp70 expression in *Drosophila melanogaster*. *Molecular and Cell Biology*, **14**, 3646-3659.
- DERISI, J. L., VASHWANATH, R. L. and BROWN, P., 1997, Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680-686.
- DIATCHENKO, L., LAU, Y.-F. C., CAMPBELL, A. P., CHENCHIK, A., MOQADAM, F., HUANG, B., LUKYANOV, K., GURSKAYA, N., SVERDLOV, E. D. and SIEBERT, P. D., 1996, Suppression subtractive hybridisation: A method for generating differentially regulated or tissue-specific cDNA probes and libraries. *Proceedings of the National Academy of Sciences (USA)*, **93**, 6025-6030.
- DOGRA, S. C., WHITELAW, M. L. and MAY, B. K., 1998, Transcriptional activation of cytochrome P450 genes by different classes of chemical inducers. *Clinical and Experimental Pharmacology and Physiology*, **25**, 1-9.
- DUGUID, J. R. and DINAUER, M. C., 1990, Library subtraction of *in vitro* cDNA libraries to identify differentially expressed genes in scrapie infection. *Nucleic Acids Research*, **18**, 2789-2792.
- DUNBAR, P. R., OGG, G. S., CHEN, J., RUST, N., VAN DER BRUGGEN, P. and CERUNDOLO, V., 1998, Direct isolation, phenotyping and cloning of low-frequency antigen-specific cytotoxic T lymphocytes from peripheral blood. *Current Biology*, **26**, 413-416.

- FITZPATRICK, D. R., GERMAIN -LEE, E. and VALLE, D., 1995, Isolation and characterisation of rat and human cDNAs encoding a novel putative peroxisomal enoyl-CoA hydratase. *Genomics*, **27**, 457-466.
- FOSS, D. L., BAARSCH, M. J. and MURTAUGH, M. P., 1998, Regulation of hypoxanthine phosphoribosyltransferase, glyceraldehyde-3-phosphate dehydrogenase and beta-actin mRNA expression in porcine immune cells and tissues. *Animal Biotechnology*, **9**, 67-78.
- FRYE, R. A., BENZ, C. C. and LIU, E., 1989, Detection of amplified oncogenes by differential polymerase chain reaction. *Oncogene*, **4**, 1153-1157.
- GEISINGER, A., RODRIGUEZ, R., ROMERO, V. and WETTSTEIN, R., 1997, A simple method for screening cDNAs arising from the cloning of RNA differential display bands. *Elsevier Trends Journals Technical Tips Online*, <http://tto.trends.com>, document T01110.
- GRESS, T. M., HOEISEL, J. D., LENNON, G. G., ZEHETNER, G. and LEHRACH, H., 1992, Hybridisation fingerprinting of high density cDNA filter arrays with cDNA pools derived from whole tissues. *Mammalian Genome*, **3**, 609-619.
- GRIFFIN, G. and KRISHNA, S., 1998, Cytokines in infectious diseases. *Journal of the Royal College of Physicians, London*, **32**, 195-198.
- GROENINK, M. and LEEGWATER, A. C. J., 1996, Isolation of delayed early genes associated with liver regeneration using Clontech PCR-select subtraction technique. *Clontechniques*, **XI**, 23-24.
- GUIMARAES, M. J., BAZAN, J. F., ZLOTNIK, A., WILES, M. V., GRIMALDI, J. C., LEE, F. and MCCLANAHAN, T., 1995b, A new approach to the study of haematopoietic development in the yolk sac and embryoid bodies. *Development*, **121**, 3335-3346.
- GUIMARAES, M. J., LEE, F., ZLOTNIK, A. and MCCLANAHAN, T., 1995a, Differential display by PCR: novel findings and applications. *Nucleic Acids Research*, **23**, 1832-1833.
- GURSKAYA, N. G., DIATCHENKO, L., CHENCHIK, P. D., SIEBERT, P. D., KHASPEKOV, G. L., LUKYANOV, K. A., VAGNER, L. L., ERMOLAEVA, O. D., LUKYANOV, S. A. and SVERDLOV, E. D., 1996, Equalising cDNA subtraction based on selective suppression of polymerase chain reaction: Cloning of Jurkat cell transcripts induced by phytohemagglutinin and phorbol 12-Myristate 13-Acetate. *Analytical Biochemistry*, **240**, 90-97.
- HAMPSON, I. N. and HAMPSON, L., 1997, CCLS and DROP—subtractive cloning made easy. *Life Science News* (A publication of Amersham Life Science), **23**, 22-24.
- HAMPSON, I. N., HAMPSON, L. and DEXTER, T. M., 1996, Directional random oligonucleotide primed (DROP) global amplification of cDNA: its application to subtractive cDNA cloning. *Nucleic Acids Research*, **24**, 4832-4835.
- HAMPSON, I. N., POPE, L., COWLING, G. J. and DEXTER, T. M., 1992, Chemical cross linking subtraction (CCLS): a new method for the generation of subtractive hybridisation probes. *Nucleic Acids Research*, **20**, 2899.
- HARA, E., KATO, T., NAKADA, S., SEKIYA, S. and ODA, K., 1991, Subtractive cDNA cloning using oligo(dT)30-latex and PCR: isolation of cDNA clones specific to undifferentiated human embryonal carcinoma cells. *Nucleic Acids Research*, **19**, 7097-7104.
- HATADA, I., HAYASHIZAKE, Y., HIROTSUNE, S., KOMATSUBARA, H. and MUKAI, T., 1991, A genomic scanning method for higher organisms using restriction sites as landmarks. *Proceedings of the National Academy of Sciences (USA)*, **88**, 9523-9527.
- HECHT, N., 1998, Molecular mechanisms of male sperm cell differentiation. *Bioessays*, **20**, 555-561.
- HEDRICK, S., COHEN, D. I., NIELSEN, E. A. and DAVIS, M. E., 1984, Isolation of T cell-specific membrane-associated proteins. *Nature*, **308**, 149-153.
- HERTZ, R., SECKBACH, M., ZAKIN, M. M. and BAR-TANA, J., 1996, Transcriptional suppression of the transferrin gene by hypolipidemic peroxisome proliferators. *Journal of Biological Chemistry*, **271**, 218-224.
- HEUVAL, J. P. V., CLARK, G. C., KOHN, M. C., TRITSCHER, A. M., GREENLEE, W. F., LUCIER, G. W. and BELL, D. A., 1994, Dioxin-responsive genes: Examination of dose-response relationships using quantitative reverse transcriptase-polymerase chain reaction. *Cancer Research*, **54**, 62-68.
- HILLIER, L. D., LENNON, G., BECKER, M., BONALDO, M. F., CHIAPPELLI, B., CHISSOE, S., DIETRICH, N., DUBUQUE, T., FAVELLO, A., GISH, W., HAWKINS, M., HULTMAN, M., KUCABA, T., LACY, M., LE, M., LE, N., MARDIS, E., MOORE, B., MORRIS, M., PARSONS, J., PRANGE, C., RIFKIN, L., ROHLFING, T., SCHELLENBERG, K., SOARES, M. B., TAN, F., THIERRY-MEG, J., TREVASKIS, E., UNDERWOOD, K., WOHLDMAN, P., WATERSTON, R., WILSON, R. and MARRA, M., 1996, Generation and analysis of 280,000 human expressed sequence tags. *Genome Research*, **6**, 807-828.
- HUBANK, M. and SCHATZ, D. G., 1994, Identifying differences in mRNA expression by representational difference analysis. *Nucleic Acids Research*, **22**, 5640-5648.
- HUNTER, T., 1991, Cooperation between oncogenes. *Cell*, **64**, 249-270.
- IVANOVA, N. B. and BELYAVSKY, A. V., 1995, Identification of differentially expressed genes by restriction endonuclease-based gene expression fingerprinting. *Nucleic Acids Research*, **23**, 2954-2958.
- JAMES, B. D. and HIGGINS, S. J., 1985, *Nucleic Acid Hybridisation* (Oxford: IRL Press Ltd).
- KAS-DEELEN, A. M., HARMSSEN, M. C., DE MAAR, E. F. and VAN SON, W. J., 1998, A sensitive method for

- quantifying cytomegalic endothelial cells in peripheral blood from cytomegalovirus-infected patients. *Clinical Diagnostic and Laboratory Immunology*, 5, 622-626.
- KILTY, I. and VICKERS, P., 1997, Fractionating DNA fragments generated by differential display PCR. *Strategies Newsletter* (Stratagene), 10, 50-51.
- KLEINJAN, D.-J. and VAN HEYNINGEN, V., 1998, Position effect in human genetic disease. *Human and Molecular Genetics*, 7, 1611-1618.
- KO, M. S., 1990, An 'equalized cDNA library' by the reassociation of short double-stranded cDNAs. *Nucleic Acids Research*, 18, 5705-5711.
- LAKE, B. G., EVANS, J. G., CUNNINGHAME, M. E. and PRICE, R. J., 1993, Comparison of the hepatic effects of Wy-14,643 on peroxisome proliferation and cell replication in the rat and Syrian hamster. *Environmental Health Perspectives*, 101, 241-248.
- LAKE, B. G., EVANS, J. G., GRAY, T. J. B., KOROSI, S. A. and NORTH, C. J., 1989, Comparative studies of nafenopin-induced hepatic peroxisome proliferation in the rat, Syrian hamster, guinea pig and marmoset. *Toxicology and Applied Pharmacology*, 99, 148-160.
- LENNARD, M. S., 1993, Genetically determined adverse drug reactions involving metabolism. *Drug Safety*, 9, 60-77.
- LEVY, S., TODD, S. C. and MAECKER, H. T., 1998, CD81(TAPA-1): a molecule involved in signal transduction and cell adhesion in the immune system. *Annual Review of Immunology*, 16, 89-109.
- LIANG, P. and PARDEE, A. B., 1992, Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science*, 257, 967-971.
- LIANG, P., AVERBOUKH, L., KEYOMARSI, K., SAGER, R. and PARDEE, A., 1992, Differential display and cloning of messenger RNAs from human breast cancer versus mammary epithelial cells. *Cancer Research*, 52, 6966-6968.
- LIANG, P., AVERBOUKH, L. and PARDEE, A. B., 1993, Distribution & cloning of eukaryotic mRNAs by means of differential display refinements and optimisation. *Nucleic Acids Research*, 21, 3269-3275.
- LIANG, P., BAUER, D., AVERBOUKH, L., WARTHOF, P., ROHRWILD, M., MULLER, H., STRAUSS, M. and PARDEE, A. B., 1995, Analysis of altered gene expression by differential display. *Methods in Enzymology*, 254, 304-321.
- LINSKENS, M. H., FENG, J., ANDREWS, W. H., ENLOW, B. E., SAATI, S. M., TONKIN, L. A., FUNK, W. D. and VILLEPONTEAU, B., 1995, Cataloging altered gene expression in young and senescent cells using enhanced differential display. *Nucleic Acids Research*, 23, 3244-3251.
- LISITSYN, N., LISITSYN, N. and WIGLER, M., 1993, Cloning the differences between two complex genomes. *Science*, 259, 946-951.
- LOHMANN, J., SCHICKLE, H. and BOSCH, T. C. G., 1995, REN Display, a rapid and efficient method for non-radioactive differential display and mRNA isolation. *Biotechniques*, 18, 200-202.
- LUNNEY, J. K., 1998, Cytokines orchestrating the immune response. *Reviews in Science and Technology*, 17, 84-94.
- MAKOWSKA, J. M., GIBSON, G. G. and BONNER, F. W., 1992, Species differences in ciprofibrate-induction of hepatic cytochrome P450A1 and peroxisome proliferation. *Journal of Biochemical Toxicology*, 7, 183-191.
- MALDARELLI, F., XIANG, C., CHAMOUN, G. and ZEICHNER, S. L., 1998, The expression of the essential nuclear splicing factor SC35 is altered by human immunodeficiency virus infection. *Virus Research*, 53, 39-51.
- MATHIEU-DAUDE, F., CHENG, R., WELSH, J. and MCCLELLAND, M., 1996, Screening of differentially amplified cDNA products from RNA arbitrarily primed PCR fingerprints using single strand conformation polymorphism (SSCP) gels. *Nucleic Acids Research*, 24, 1504-1507.
- McKENZIE, D. and DRAKE, D., 1997, Identification of differentially expressed gene products with the castaway system. *Strategies Newsletter* (Stratagene), 10, 19-20.
- MCCLELLAND, M., MATHIEU-DAUDE, F. and WELSH, J., 1996, RNA fingerprinting and differential display using arbitrarily primed PCR. *Trends in Genetics*, 11, 242-246.
- MECHLER, B. and RABBITTS, T. H., 1981, Membrane-bound ribosomes of myeloma cells. IV. mRNA complexity of free and membrane-bound polysomes. *Journal of Cell Biology*, 88, 29-36.
- MEYER, U. A. and ZANGER, U. M., 1997, Molecular mechanisms of genetic polymorphisms of drug metabolism. *Annual Review of Pharmacology and Toxicology*, 37, 269-296.
- MOHLER, K. M. and BUTLER, L. D., 1991, Quantitation of cytokine mRNA levels utilizing the reverse transcriptase-polymerase chain reaction following primary antigen-specific sensitization in vivo—I. Verification of linearity, reproducibility and specificity. *Molecular Immunology*, 28, 437-447.
- MURPHY, L. D., HERZOG, C. E., RUDICK, J. B., TITO FOJO, A. and BATES, S. E., 1990, Use of the polymerase chain reaction in the quantitation of the *mdr-1* gene expression. *Biochemistry*, 29, 10351-10356.
- NELSON, D. R., KOYMANS, L., KAMATAKI, T., STEGEMAN, J. J., FEYEREISEN, R., WAXMAN, D. J., WATERMAN, M. R., GOTOH, O., COON, M. J., ESTABROOK, R. W., GUNSALUS, I. C. and NEBERT, D. W., 1996, Update on new sequences, gene mapping, accession numbers and nomenclature. *Pharmacogenetics*, 6, 1-42.

- NISHIO, Y., AIELLO, L. P. and KING, G. L., 1994, Glucose induced genes in bovine aortic smooth muscle cells identified by mRNA differential display. *FASEB Journal*, **8**, 103-106.
- O'NEILL, M. J. and SINCLAIR, A. H., 1997, Isolation of rare transcripts by representational difference analysis. *Nucleic Acids Research*, **25**, 2681-2682.
- ORTON, T. C., ADAM, H. K., BENTLEY, M., HOLLOWAY, B. and TUCKER, M. J., 1984, Clobuzarit: species differences in the morphological and biochemical response of the liver following chronic administration. *Toxicology and Applied Pharmacology*, **73**, 138-151.
- PELKONEN, O., MAENPAA, J., TAAVITSAINEN, P., RAUTIO, A. and RAUNIO, H., 1998, Inhibition and Induction of human cytochrome P450 (CYP) enzymes. *Xenobiotica*, **28**, 1203-1253.
- PHILIPS, S. M., BENDALL, A. J. and RAMSHAW, I. A., 1990, Isolation of genes associated with high metastatic potential in rat mammary adenocarcinomas. *Journal of the National Cancer Institute*, **82**, 199-203.
- PRASHAR, Y. and WEISSMAN, S. M., 1996, Analysis of differential gene expression by display of 3' end restriction fragments of cDNAs. *Proceedings of the National Academy of Sciences (USA)*, **93**, 659-663.
- RAGNO, S., ESTRADA, I., BUTLER, R. and COLSTON, M. J., 1997, Regulation of macrophage gene expression following invasion by *Mycobacterium tuberculosis*. *Immunology Letters*, **57**, 143-146.
- RAMANA, K. V. and KOHLI, K. K., 1998, Gene regulation of cytochrome P450—an overview. *Indian Journal of Experimental Biology*, **36**, 437-446.
- RICHARD, L., VELASCO, P. and DETMAR, M., 1998, A simple immunomagnetic protocol for the selective isolation and long-term culture of human dermal microvascular endothelial cells. *Experimental Cell Research*, **240**, 1-6.
- ROCKETT, J. C., ESDAILE, D. J. and GIBSON, G. G., 1997, Molecular profiling of non-genotoxic hepatocarcinogenesis using differential display reverse transcription-polymerase chain reaction (ddRT-PCR). *European Journal of Drug Metabolism and Pharmacokinetics*, **22**, 329-333.
- RODRICKS, J. V. and TURNBULL, D., 1987, Inter-species differences in peroxisomes and peroxisome proliferation. *Toxicology and Industrial Health*, **3**, 197-212.
- ROGLER, G., HAUSMANN, M., VOGL, D., ASCHENBRENNER, E., ANDUS, T., FALK, W., ANDRESEN, R., SCHOLMERICH, J. and GROSS, V., 1998, Isolation and phenotypic characterization of colonic macrophages. *Clinical and Experimental Immunology*, **112**, 205-215.
- ROHN, W. M., LEE, Y. J. and BENVENISTE, E. N., 1996, Regulation of class II MHC expression. *Critical Reviews in Immunology*, **16**, 311-330.
- RUDIN, C. M. and THOMPSON, C. B., 1998, B-cell development and maturation. *Seminars in Oncology*, **25**, 435-446.
- SAKAGUCHI, N., BERGER, C. N. and MELCHERS, F., 1986, Isolation of a cDNA copy of an RNA species expressed in murine pre-B cells. *EMBO Journal*, **5**, 2139-2147.
- SAMBROOK, J., FRITSCH, E. F. and MANIATIS, T., 1989, Gel electrophoresis of DNA. In N. Ford, M. Nolan and M. Ferguson (eds), *Molecular Cloning—A laboratory manual*, 2nd edition (New York: Cold Spring Harbour Laboratory Press), Volume 1, pp. 6-37.
- SARGENT, T. D. and DAWID, I. B., 1983, Differential gene expression in the gastrula of *Xenopus laevis*. *Science*, **222**, 135-139.
- SCHENA, M., SHALON, D., HELLER, R., CHAI, A., BROWN, P. O. and DAVIS, R. W., 1996, Parallel human genome analysis: Microarray-based expression monitoring of 1000 genes. *Proceedings of the National Academy of Sciences (USA)*, **93**, 10614-10619.
- SCHNEIDER, C., KING, R. M. and PHILIPSON, L., 1988, Genes specifically expressed at growth arrest of mammalian cells. *Cell*, **54**, 787-793.
- SCHNEIDER-MAUNOURY, S., GILARDI-HEBENSTREIT, P. and CHARNAY, P., 1998, How to build a vertebrate hindbrain. Lessons from genetics. *C R Academy of Science III*, **321**, 819-834.
- SEMENZA, G. L., 1994, Transcriptional regulation of gene expression: mechanisms and pathophysiology. *Human Mutations*, **3**, 180-199.
- SEWALL, C. H., BELL, D. A., CLARK, G. C., TRITSCHER, A. M., TULLY, D. B., VANDEN HEUVEL, J. and LUCIER, G. W., 1995, Induced gene transcription: implications for biomarkers. *Clinical Chemistry*, **41**, 1829-1834.
- SINGH, N., AGRAWAL, S. and RASTOGI, A. K., 1997, Infectious diseases and immunity: special reference to major histocompatibility complex. *Emerging Infectious Diseases*, **3**, 41-49.
- SMITH, N. R., LI, A., ALDERSLEY, M., HIGH, A. S., MARKHAM, A. F. and ROBINSON, P. A., 1997, Rapid determination of the complexity of cDNA bands extracted from DDRT-PCR polyacrylamide gels. *Nucleic Acids Research*, **25**, 3552-3554.
- SOMPAYRAC, L., JANE, S., BURN, T. C., TENEN, D. G. and DANNA, K. J., 1995, Overcoming limitations of the mRNA differential display technique. *Nucleic Acids Research*, **23**, 4738-4739.
- ST JOHN, T. P. and DAVIS, R. W., 1979, Isolation of galactose-inducible DNA sequences from *Saccharomyces cerevisiae* by differential plaque filter hybridisation. *Cell*, **16**, 443-452.
- SUN, Y., HEGAMER, G. and COLBURN, N. H., 1994, Molecular cloning of five messenger RNAs differentially expressed in preneoplastic or neoplastic JB6 mouse epidermal cells: one is homologous to human tissue inhibitor of metalloproteinases-3. *Cancer Research*, **54**, 1139-1144.

- SUNG, Y. J. and DENMAN, R. B., 1997, Use of two reverse transcriptases eliminates false-positive results in differential display. *Biotechniques*, **23**, 462-464.
- SUTTON, G., WHITE, O., ADAMS, M. and KERLAVAGE, A., 1995, TIGR Assembler; A new tool for assembling large shotgun sequencing projects. *Genome Science and Technology*, **1**, 9-19.
- SUZUKI, Y., SEKIYA, T. and HAYASHI, K., 1991, Allele-specific polymerase chain reaction: a method for amplification and sequence determination of a single component among a mixture of sequence variants. *Analytical Biochemistry*, **192**, 82-84.
- SYED, V., GU, W. and HECHT, N. B., 1997, Sertoli cells in culture and mRNA differential display provide a sensitive early warning assay system to detect changes induced by xenobiotics. *Journal of Andrology*, **18**, 264-273.
- UITTERLINDEN, A. G., SLAGBOOM, P., KNOOK, D. L. and VIJGL, J., 1989, Two-dimensional DNA fingerprinting of human individuals. *Proceedings of the National Academy of Sciences (USA)*, **86**, 2742-2746.
- ULLMAN, K. S., NORTHROP, J. P., VERWEIJ, C. L. and CRABTREE, G. R., 1990, Transmission of signals from the T lymphocyte antigen receptor to the genes responsible for cell proliferation and immune function: the missing link. *Annual Review of Immunology*, **8**, 421-452.
- VASMATZIS, G., ESSAND, M., BRINKMANN, U., LEE, B. and PASTON, I., 1998, Discovery of three genes specifically expressed in human prostate by expressed sequence tag database analysis. *Proceedings of the National Academy of Sciences (USA)*, **95**, 300-304.
- VELCULESCU, V. E., ZHANG, L., VOGELSTEIN, B. and KINZLER, K. W., 1995, Serial analysis of gene expression. *Science*, **270**, 484-487.
- VOELTZ, G. K. and STEITZ, J. A., 1998, AuuuA sequences direct mRNA deadenylation uncoupled from decay during *Xenopus* early development. *Molecular and Cell Biology*, **18**, 7537-7545.
- VOGELSTEIN, B. and KINZLER, K. W., 1993, The multistep nature of cancer. *Trends in Genetics*, **9**, 138-141.
- WALTER, J., BELFIELD, M., HAMPSON, I. and READ, C., 1997, A novel approach for generating subtractive probes for differential screening by CCLS. *Life Science News*, **21**, 13-14.
- WAN, J. S., SHARP, S. J., POIRIER, G. M.-C., WAGAMAN, P. C., CHAMBERS, J., PYATI, J., HOM, Y.-L., GALINDO, J. E., HUVAR, A., PETERSON, P. A., JACKSON, M. R. and ERLANDER, M. G., 1996, Cloning differentially expressed mRNAs. *Nature Biotechnology*, **14**, 1685-1691.
- WALTER, J., BELFIELD, M., HAMPSON, I. and READ, C., 1997, A novel approach for generating subtractive probes for differential screening by CCLS. *Life Science News*, **21**, 13-14.
- WANG, Z. and BROWN, D. D., 1991, A gene expression screen. *Proceedings of the National Academy of Sciences (USA)*, **88**, 11505-11509.
- WAWER, C., RUGGEBERG, H., MEYER, G. and MUYZER, G., 1995, A simple and rapid electrophoresis method to detect sequence variation in PCR-amplified DNA fragments. *Nucleic Acids Research*, **23**, 4928-4929.
- WELSH, J., CHADA, K., DALAL, S. S., CHENG, R., RALPH, D. and MCCLELLAND, M., 1992, Arbitrarily primed PCR fingerprinting of RNA. *Nucleic Acids Research*, **20**, 4965-4970.
- WONG, H., ANDERSON, W. D., CHENG, T. and RIABOWOL, K. T., 1994, Monitoring mRNA expression by polymerase chain reaction: the 'primer-dropping' method. *Analytical Biochemistry*, **223**, 251-258.
- WONG, K. K. and MCCLELLAND, M., 1994, Stress-inducible gene of *Salmonella typhimurium* identified by arbitrarily primed PCR of RNA. *Proceedings of the National Academy of Sciences (USA)*, **91**, 639-643.
- WYNFORD-THOMAS, D., 1991, Oncogenes and anti-oncogenes; the molecular basis of tumour behaviour. *Journal of Pathology*, **165**, 187-201.
- XHU, D., CHAN, W. L., LEUNG, B. P., HUANG, F. P., WHEELER, R., PIEDRAFITA, D., ROBINSON, J. H. and LIEW, F. Y., 1998, Selective expression of a stable cell surface molecule on type 2 but not type 1 helper T cells. *Journal of Experimental Medicine*, **187**, 787-794.
- YANG, M. and SYTOWSKI, A. J., 1996, Cloning differentially expressed genes by linker capture subtraction. *Analytical Biochemistry*, **237**, 109-114.
- ZHAO, N., HASHIDA, H., TAKAHASHI, N., MISUMI, Y. and SAKAKI, Y., 1995, High-density cDNA filter analysis: a novel approach for large scale quantitative analysis of gene expression. *Gene*, **156**, 207-213.
- ZHAO, X. J., NEWSOME, J. T. and CIHLAR, R. L., 1998, Up-regulation of two *Candida albicans* genes in the rat model of oral candidiasis detected by differential display. *Microbial Pathogenesis*, **25**, 121-129.
- ZIMMERMANN, C. R., ORR, W. C., LECLERC, R. F., BARNARD, C. and TIMBERLAKE, W. E., 1980, Molecular cloning and selection of genes regulated in *Aspergillus* development. *Cell*, **21**, 709-715.

Molecular profiling of non-genotoxic hepatocarcinogenesis using differential display reverse transcription-polymerase chain reaction (ddRT-PCR)

J.C. ROCKETT¹, D.J. ESDAILE² and G.G. GIBSON¹

¹Molecular Toxicology Group, School of Biological Sciences, University of Surrey, Guildford, UK

²Rhône-Poulenc Agrochemicals, Sophia Antipolis, France

Keywords : ddRT-PCR, non-genotoxic hepatocarcinogenesis, phenobarbital, rat, WY-14,643

SUMMARY

The technique of differential display reverse transcription-polymerase chain reaction (ddRT-PCR) has been used to produce unique profiles of up-regulated and down-regulated gene expression in the liver of male Wistar rats following short term exposure to the non-genotoxic hepatocarcinogens, phenobarbital and WY-14,643. Animals were treated for 3 days, whereupon their livers were extracted and snap frozen. mRNA was prepared from the livers and used for ddRT-PCR. Individual bands from the differential displays were extracted and cloned. False positives were eliminated by dotblot screening and true positives then sequenced and identified.

INTRODUCTION

Safety evaluation of new chemicals usually necessitates the examination of genotoxic and carcinogenic potential using short-term in vitro and in vivo genotoxicity assays augmented by chronic bioassay tests. The short-term assays have proved useful in the early identification of potential genotoxic carcinogens, but their value is limited by observations which suggest that approximately 60% of chemicals identified as carcinogens in life-exposure studies produce mainly negative findings in short-term genotoxicity tests (1,2). Thus, there is currently no reliable and rapid means of evaluating the carcinogenic risk of new chemicals which fall into this latter group of compounds, termed non-genotoxic (or epigenetic) carcinogens.

It is now evident that non-genotoxic carcinogens constitute a group of chemicals which are not only divergent in their interspecies toxicity, but also demonstrate different target organ selectivities and mechanisms of action (3,4). Elucidation of the molecular mechanisms underlying non-genotoxic carcinogenesis is currently underway, but the picture is still far from complete. It is anticipated that a better understanding of the early changes in genetic expression following exposure to non-genotoxic carcinogens will aid development of experimental strategies to identify cellular markers which are diagnostic for this type of toxicity.

Subtractive ddRT-PCR is a recently developed technique which facilitates the preferential amplification of gene products that demonstrate altered expression in target tissue(s) following exposure to chemical stimuli. Furthermore, using this technique, no prior knowledge of the specific genes which are up/down regulated is required. In the current study, we have undertaken to develop a specific and rapid assay for non-genotoxic carcinogens using the technique of ddRT-PCR. This has allowed us to identify characteristic

Please send reprint requests to : Dr John Rockett, Molecular Toxicology Group, School of Biological Sciences, University of Surrey, Guildford, Surrey GU2 5XH, UK.

patterns of gene regulation following administration of two different non-genotoxic carcinogens (phenobarbital and Wy-14,643) and the subsequent identification of individual gene species which are regulated by this xenobiotic treatment.

MATERIALS AND METHODS

Animals and treatment

Phenobarbital (BDH, Poole, UK; 100 mg/kg/day) or [4-chloro-6-(2,3-xylydino)-2-pyrimidinylthio] acetic acid (Wy-14,643) (Campo, Emmerich; 250 mg/kg/day) was administered by gavage to groups of 3 male Wistar rats (150–200 g) on three consecutive days, whilst control animals received nothing. All animals had free access to food (rat and mouse standard diet, B&K Universal, Hull, UK) and water. The animals were killed on the fourth day, whereupon their livers were excised, sliced into 0.5 cm cubes, snap frozen in liquid nitrogen and then stored at -70°C .

mRNA extraction

Up to 0.25 g of each frozen liver sample was ground under liquid nitrogen using a mortar and pestle. mRNA was extracted from the ground liver using Promega's PolyAtract® System 1000 (Promega, Madison, WI, USA) according to the technical manual. The mRNA was DNase-treated (Promega, final concentration 10 U/ml) before phenol/chloroform extraction and ethanol precipitation. The mRNA was resuspended at a final concentration 500–1000 ng/ μl .

ddRT-PCR

This was carried out using the PCR-Select™ cDNA Subtraction Kit (Clontech, Palo Alto, CA, USA) according to the manufacturer's instructions. Final PCR reactions were run on a 2% Metaphor agarose (FMC, Rockland, MD, USA) gel containing ethidium bromide (Sigma, Dorset, UK) and then overstained for 30 min with SYBR Green I DNA stain (FMC, 1:10 000 dilution in TAE).

Band extraction and cloning

Each discernible band from the differential display pattern was extracted from the gel with a scalpel and

the DNA eluted using a Genelute™ Agarose Spin Column (Supelco, Bellefonte). An aliquot of the eluted DNA (5 μl) was re-amplified using the original ddRT-PCR nested primers and electrophoresed on a 2% agarose gel. The re-amplified band was extracted from the gel (as above) and the eluted DNA ligated directly into the TOPO TA Cloning® vector (Invitrogen, Carlsbad) before transformation in *Escherichia coli* TOP10F' One Shot™ cells (Invitrogen).

Stage 1 screening

Twelve transformed (white) colonies from each band were grown up for 6 h in 200 μl LB broth containing ampicillin (Sigma, 50 $\mu\text{g}/\text{ml}$) and 1 μl of this amplified by PCR reaction (as specified in ddRT-PCR technical manual). One quarter of the completed reaction was electrophoresed on a standard 2% agarose gel and one quarter on a 2% agarose gel containing HA Yellow (Hanse Analytik GmbH, Bremen, Germany, 1 U/ μl) to discern the different cloning products. The remainder was used to prepare duplicate dotblots on Hybond N+ (nylon) membranes (Amersham, Little Chalfont, UK). Cultures containing different cloning products were grown up and a plasmid miniprep prepared from each (Wizard Plus SV Minipreps DNA Purification System, Promega) according to the manufacturer's instructions.

Stage II screening

The duplicate dotblots were probed with: (a) the final differential display reaction; and (b) the 'reverse-subtracted' differential display reaction. To make the 'reverse-subtracted' probe, the subtractive hybridisation step of the ddRT-PCR procedure was carried out using the original tester cDNA as a driver and the driver as a tester. Probing and visualisation were carried out using the ECL Direct Nucleic Acid Labelling and Detection System (Amersham) according to the manufacturer's instructions. Those clones which were positive for (a) but negative for (b), or showed a substantially larger positive signal with (a) compared to (b), were chosen for further analysis.

DNA sequencing

Positive clones as identified above were sequenced on an automated ABI DNA sequencer (Applied Biosystems, Warrington, UK).

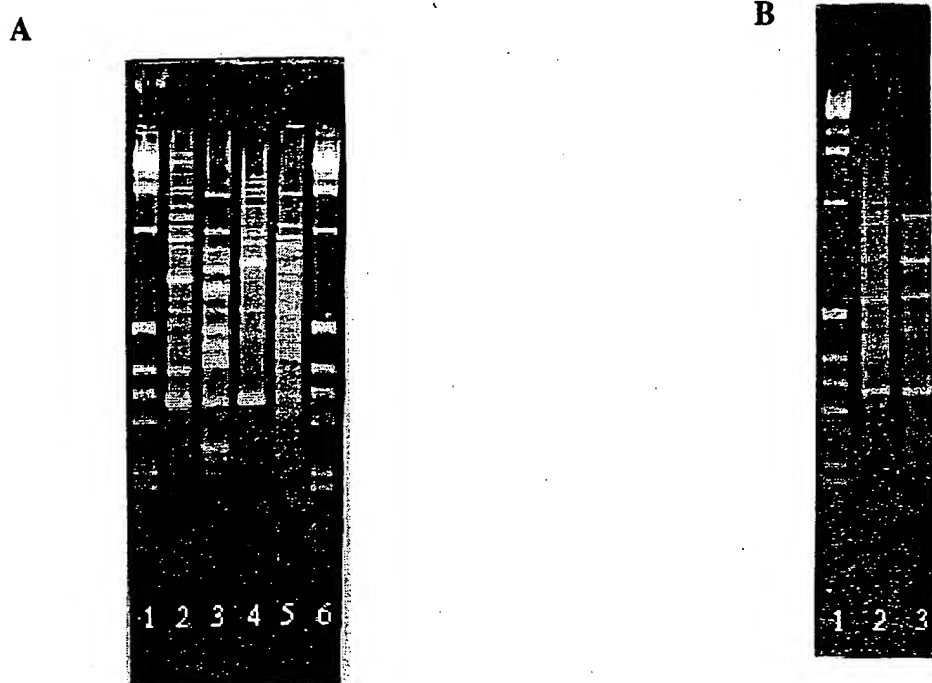


Fig. 1 : (A) Subtractive ddRT-PCR patterns obtained from rat liver following 3-day treatment with WY-14,643 or phenobarbital. Lane 1, 1 kb ladder; lane 2, genes up-regulated following Wy,14-643 treatment; lane 3, genes down-regulated following Wy,14-643 treatment; lane 4, genes up-regulated following phenobarbital treatment; lane 5, genes down-regulated following phenobarbital treatment; and lane 6, 1kb ladder. (B) Subtractive ddRT-PCR patterns obtained from rat liver showing relative changes when phenobarbital treated mRNA is subtracted from Wy-14,643-treated mRNA and vice-versa. Lane 1, 1 kb ladder; lane 2, genes showing increased expression following Wy-14,643 treatment compared to phenobarbital treatment; lane 3, genes showing increased expression following phenobarbital treatment compared to Wy-14,643 treatment. See Materials and Methods for further details.

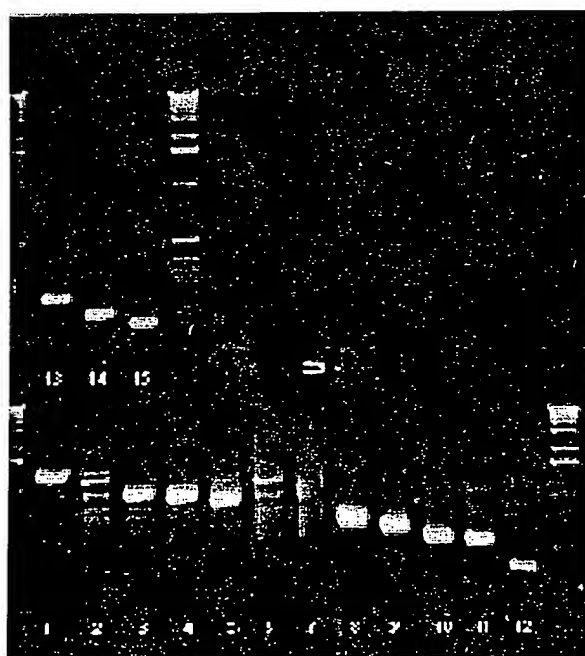


Fig. 2 : Re-amplified ddRT-PCR products which were down-regulated following phenobarbital treatment (upregulated bands were also re-amplified but gel not shown). Individual DNA bands excised from gel of ddRTR-PCR reactions were extracted, re-amplified and run on agarose gels to confirm amplification of correct band (numbered). See Materials and Methods for further details.

Table I : Rat liver genes down-regulated by phenobarbital treatment

Band number (Fig. 2) (Approximate size in bp)	Phenobarbital down-regulated	
	Highest sequence homology	FASTA-EMBL gene identification
1 (1500)	95.3%	Rat mRNA for 3-oxoacyl-CoA thiolase
2 (1200)	92.3%	Rat hemopoxin mRNA
3 (1000)	91.7%	<i>R. rattus</i> alpha-2u-globulin mRNA
7 (700)	Clone 1	77.2% <i>M. musculus</i> mRNA for CI inhibitor
	Clone 2	94.5% Rat electron transfer flavoprotein
	Clone 3	91.0% Mouse topoisomerase 1 (Topo 1) mRNA
8 (650)	Clone 1	86.9% Soares 2NbMT <i>M. musculus</i> (EST)
	Clone 2	96.2% Rat alpha-2u-globulin (s-type) mRNA
9 (600)	Clone 1	86.9% Soares mouse NML <i>M. musculus</i> (EST)
	Clone 2	82.0% Soares p3NMF19.5 <i>M. musculus</i> (EST)
10 (550)	73.8%	Soares mouse NML <i>M. musculus</i> (EST)
11 (525)	95.7%	NCI_CGAP_Pr1 <i>H. sapiens</i> (EST)
12 (375)	100.0%	<i>R. norvegicus</i> mRNA for ribosomal protein
13 (230)	Clone 1	97.2% Soares mouse embryo NbME135 (EST)
	Clone 2	100.0% Rat fibrinogen B-beta-chain
	Clone 3	100.0% Rat apolipoprotein E gene
14 (170)	96.0%	Soares p3NMF19.5 <i>M. musculus</i> (EST)
15 (140)	97.3%	Stratagene mouse testis (EST)
Others: (300)	96.7%	<i>R. norvegicus</i> RASP 1 mRNA
(275)	93.1%	Soares mouse mammary gland (EST)

EST = expressed sequence tag.

Bands 4-6 were shown to be false positives by dotblot analysis and, therefore, not sequenced.

Table II : Rat liver genes up-regulated by phenobarbital treatment

Band number (Approximate size in bp)	Phenobarbital up-regulated	
	Highest sequence homology	FASTA-EMBL gene identification
5 (1300)	93.5%	Rat cytochrome P450IIB1
7 (1000)	95.1%	mRNA for rat preproalbumin
		Rat serum albumin mRNA
8 (950)	98.3%	NCI_CGAP_Pr1 <i>H. sapiens</i> (EST)
10 (850)	95.7%	Rat cytochrome P450IIB1
11 (800)	Clone 1	94.9% Rat cytochrome P450IIB1
	Clone 2	75.3% Rat cytochrome p450-L (p450IIB2)
12 (750)	93.8%	Rat TRPM-2 mRNA
		Rat mRNA for sulfated glycoprotein
15 (600)	92.9%	mRNA for rat preproalbumin
		Rat serum albumin mRNA
16 (550)	Clone 1	95.2% Rat cytochrome P450IIB1
	Clone 2	93.6% Rat haptoglobulin mRNA partial alpha
21 (350)	99.3%	<i>R. norvegicus</i> genes for 18S, 5.8S & 28S rRNA

EST = expressed sequence tag.

Bands 1-4, 6, 9, 13, 14 and 17-20 shown to be false positives by dotblot analysis and, therefore, not sequenced.

Identification of differentially-regulated genes

Gene-sequences were identified using the FASTA programme (<http://www.ebi.ac.uk/htbin/fasta.py?request>) to search all EMBL databases for matching DNA sequences.

RESULTS

Figure 1A,B shows the ddRT-PCR patterns of genes showing altered expression in rat liver following 3 day treatment with phenobarbital or Wy-14,643. Individual bands were isolated from the phenobarbital-modulated patterns (both up- and down-regulated), re-amplified (Fig. 2), cloned, screened for false positives and then identified. Those xenobiotic-modulated gene products identified to date are listed in Tables I and II.

DISCUSSION

The advent of combinatorial chemistry has led to the synthesis of millions of new chemical compounds, many of which may be potentially useful in pharmaceutical, agricultural or industrial applications. However, whilst there are tests available for those posing a genotoxic activity, there remains no short-term assay able to identify those chemicals which may belong to the non-genotoxic group of carcinogens.

We have used an adaptation of the subtractive hybridisation method – ddRT-PCR – to produce characteristic profiles or 'fingerprints' of those genes which are up-regulated or down-regulated in male rat liver following acute exposure to test chemicals. The ddRT-PCR profiles are characteristic and unique for each of the 2 compounds studied to date.

A number of those gene species showing altered expression following phenobarbital treatment have been cloned and identified (Tables I & II). It is interesting to note the presence of CYP2B2 in the up-regulated genes. This would, of course, be expected following exposure to phenobarbital and serves as a positive control for the method. Other genes which one might normally expect to be up-regulated do not appear in the table. However, it should be noted that not

all bands seen on the differential display were extracted and re-amplified due to their being too faint or too close to other bands to accurately excise. Furthermore, it has been well documented [(5) and references therein] that a single band extracted from a differential display often represents a composite of heterogeneous products. We are currently examining new methods to: (i) improve resolution of the differential display patterns (including 2-D agarose gels); and (ii) distinguish those ddRT-PCR products which are identical in size, but different in sequence.

Our future efforts will be directed towards determining the extent of modulation of a number of the genes reported herein using semi-quantitative RT-PCR. This should reveal the extent of changes in expression of key gene products which may be involved in non-genotoxic hepatocarcinogenesis and thus help increase understanding of this process. Furthermore, it is anticipated that aligning ddRT-PCR profiles of different non-genotoxic agents found in responsive and non-responsive species may enable identification of those genes which are mechanistically relevant to the non-genotoxic hepatocarcinogenic process. Accordingly, this approach lends itself well to the identification, characterisation and sub-classification of possible different classes of non-genotoxic carcinogens.

ACKNOWLEDGEMENT

This work was funded by Rhône-Poulenc Agrochemicals, France

REFERENCES

1. Parodi S. (1992) : Non-genotoxic factors in the carcinogenic process: problems of detection and hazard evaluation. *Toxicol. Lett.*, 64/65, 621-630.
2. Ashby J. (1992) : Prediction of non-genotoxic carcinogenesis. *Toxicol. Lett.*, 64/65, 605-612.
3. Grasso G. and Sharratt M. (1991) : Role of persistent, non-genotoxic tissue damage in rodent cancer and relevance to humans. *Annu. Rev. Pharmacol. Toxicol.*, 31, 253-287.
4. Lake B. (1995) : Mechanisms of hepatocarcinogenicity of peroxisome-proliferating drugs and chemicals. *Annu. Rev. Pharmacol. Toxicol.*, 35, 483-507.
5. Smith N.R., Li A., Aldersley M., High A.S., Markham A.E., Robinson P.A. (1997) : Rapid determination of the complexity of cDNA bands extracted from DDRT-PCR polyacrylamide gels. *Nucleic Acids Research* 25 (17). 3552-3554.

Univ. of Minn.
Bio-Medical
Library

DRUG METABOLISM KINETICS

10-11-11

PUBLISHED BY THE

Use of suppression-PCR subtractive hybridisation to identify genes that demonstrate altered expression in male rat and guinea pig livers following exposure to Wy-14,643, a peroxisome proliferator and non-genotoxic hepatocarcinogen

John C. Rockett¹, Karen E. Swales, David J. Esdaile², G. Gordon Gibson *

Molecular Toxicology Group, School of Biological Sciences, University of Surrey, Guildford, Surrey GU2 5XH, UK

Abstract

Understanding the genetic profile of a cell at all stages of normal and carcinogenic development should provide an essential aid to developing new strategies for the prevention, early detection, diagnosis and treatment of cancers. We have attempted to identify some of the genes that may be involved in peroxisome-proliferator (PP)-induced non-genotoxic hepatocarcinogenesis using suppression PCR subtractive hybridisation (SSH). Wistar rats (male) were chosen as a representative susceptible species and Duncan–Hartley guinea pigs (male) as a resistant species to the hepatocarcinogenic effects of the PP, [4-chloro-6-(2,3-xylylidino)-2-pyrimidinylthio] acetic acid (Wy-14,643). In each case, groups of four test animals were administered a single dose of Wy-14,643 (250 mg/kg per day in corn oil) by gastric intubation for 3 consecutive days. The control animals received corn oil only. On the fourth day the animals were killed and liver mRNA extracted. SSH was carried out using mRNA extracted from the rat and guinea pig livers, and used to isolate genes that were up and downregulated following Wy-14,643 treatment. These genes included some predictable (and hence positive control) species such as CYP4A1 and CYP2C11 (upregulated and downregulated in rat liver, respectively). Several genes that may be implicated in hepatocarcinogenesis have also been identified, as have some unidentified species. This work thus provides a starting point for developing a molecular profile of the early effects of a non-genotoxic carcinogen in sensitive and resistant species that could ultimately lead to a short-term assay for this type of toxicity. © 2000 Elsevier Science Ireland Ltd. All rights reserved.

Keywords: Wy-14,643; Peroxisome proliferator; Non-genotoxic hepatocarcinogenesis; Suppression PCR subtractive hybridisation; RT-PCR; Rat; Guinea pig; Gene regulation; Differential gene display; Gene profiling

* Corresponding author. Tel.: +44-1483-259704; fax: +44-1483-576978.

E-mail address: g.gibson@surrey.ac.uk (G.G. Gibson)

¹ Present address: US Environmental Protection Agency, National Health and Environmental Effects Research Laboratory, Reproductive Toxicology Section, Research Triangle Park, NC 27711, USA.

² Present address: Rhone-Poulenc Agrochemicals, Toxicology Department, Sophia-Antipolis, Nice, France.

Introduction

The advent of combinatorial chemistry and computer-aided drug design has led to a recent upsurge in the number of chemical compounds that have potential therapeutic, agricultural and industrial applications. Although it has been suggested that the contribution of synthetic chemicals to the overall incidence of human cancer is low, there still remains an absolute requirement to evaluate all new chemicals for toxic and carcinogenic potential. The latter is one of the most problematic areas of chemical safety evaluation and is usually carried out using short-term *in vitro* and *in vivo* genotoxicity assays augmented by micronucleus bioassay tests. The short-term assays have proved useful in the early identification of potential genotoxic carcinogens, but their value is limited by observations that suggest that approximately 60% of chemicals identified as carcinogens in life-exposure studies produce mainly negative findings in short-term genotoxicity tests (Ashby, 1992; Parodi, 1992). Thus, there is currently no reliable and rapid means of evaluating the carcinogenic risk of new chemicals that fall into this latter group of compounds, termed non-genotoxic (or epigenetic) carcinogens.

One approach to addressing this problem is to elucidate the molecular mechanisms by which known non-genotoxic carcinogens act. It should then be possible to identify common factors/mechanisms that can serve as early biomarkers of carcinogenic potential for new chemicals. To this end, a large number of groups have reported on the various effects of non-genotoxic compounds in various animal species (Marsman et al., 1988; Wake et al., 1993; Cattley et al., 1994; Hayashi et al., 1994; Human and Experimental Toxicology, 1994; Anderson et al., 1996). However, the mechanistic picture is still far from complete with many of those genes involved in the carcinogenic process remaining unknown, and their identification therefore remains a key goal in elucidating the molecular mechanisms by which non-genotoxic carcinogenesis occurs.

Subtractive hybridisation (SH) and related technologies such as representational difference analysis (RDA) (Hubank and Schatz, 1994) and

differential display (DD) (Liang and Pardee, 1992) can be used to aid the isolation of genes showing altered expression in target tissues following exposure to a chemical stimulus. These techniques can also be used to identify differential gene expression in neoplastic and normal cells (Liang et al., 1992), infected and normal cells (Duguid and Dinauer, 1990), differentiated and undifferentiated cells (Sargent and Dawid, 1983; Guimaraes et al., 1995), activated and dormant cells (Gurskaya et al., 1996; Wan et al., 1996), different cell types (Hedrick et al., 1984; Davis et al., 1984) amongst others. Most importantly, using such approaches, no prior knowledge of the specific genes that are upregulated/downregulated is required.

Using a variation of SH, termed suppression-PCR subtractive hybridisation (SSH) (Diatchenko et al., 1996), we have previously reported the isolation of a number of genes showing altered expression in male rat liver following acute exposure to phenobarbital (Rockett et al., 1997). In the current work we have used the same experimental approach to isolate genes that are differentially expressed in the livers of male rats and guinea pigs following short-term (3-day) exposure to the peroxisome proliferator (PP) and non-genotoxic hepatocarcinogen, Wy-14,643. We have isolated and identified a number of gene species, some of which may be important in the induction of, or protection against, non-genotoxic hepatocarcinogenesis.

2. Materials and methods

2.1. Animals and treatment

All animal experiments were undertaken in accordance with Her Majesty's Home Office Department guidelines under the auspices of approved personal and project licences. Male Wistar rats (150–200 g) and male Duncan–Hartley guinea pigs (250–300 g) were obtained from Kingman and Bantam (Hull, UK). Upon receipt, both groups were randomly assigned into two groups of four. They were maintained on a rat, mouse or guinea pig standard diet (B&K Univer-

säl, Hull) and a daily cycle of alternating 12-h periods of dark and light. The room temperature was maintained at 19°C and a relative humidity of 55%. The animals were acclimatised to this environment for 7 days before treatment commenced. [4-chloro-6-(2,3-xylydino)-2-pyrimidinylthio] acetic acid (Wy-14,643, Campo, Emmerich; 250 mg/kg per day in corn oil) was administered by gavage to the treated groups of rats and guinea pigs on 3 consecutive days, whilst control groups received an equal volume of corn oil only. During this time, all animals had free access to food and water. The animals were killed by cervical dislocation on the fourth day, and their livers immediately excised, weighed, sliced into approximately 0.5-cm cubes, snap frozen in liquid nitrogen and stored at –70°C.

2.2. mRNA extraction

Approximately 0.25 g of each frozen liver sample was ground under liquid nitrogen using a mortar and pestle. Messenger RNA was extracted from the ground liver using the PolyATtract® System 1000 kit (Promega, Madison, USA) according to the technical manual provided by the manufacturers. The mRNA was DNase-treated (RQ Rnase-free Dnase, Promega, final concentration 10 U/ml) before phenol/chloroform extraction and ethanol precipitation. The mRNA was redissolved at a final concentration 500–1000 ng/µl.

2.3. cDNA Subtraction

This was carried out using the PCR-Select™ cDNA Subtraction Kit (Clontech, Palo Alto, USA) according to the manufacturer's instructions. Subtractions were carried out with mRNAs derived from single animals. The mRNA from the remaining three animals in each group was later used for quantitative RT-PCR analysis of specific genes.

2.4. Band extraction and cloning

The secondary PCR reactions from the cDNA subtraction procedure were run on a 2%

Metaphor agarose gel (FMC, Rockland, USA) containing 0.5 µg/ml ethidium bromide (Sigma, Dorset, UK). One times TAE (0.04 M Tris-acetate, 0.001 M EDTA) was used to prepare the gel and as the running buffer. After running for 6–7 h at 3.75 V/cm, the gel was overstained for 30 min with SYBR Green I DNA stain (FMC, 1:10 000 dilution in 1 × TAE). Each discernible band of the differential display pattern was extracted from the gel with a scalpel and the DNA eluted using a Genelute™ agarose spin column (Supelco, Bellefonte, USA). Five microlitres of the eluted DNA was reamplified using the original nested (secondary) PCR primers supplied with the PCR-Select™ cDNA subtraction kit. The PCR products were electrophoresed on a 2% standard agarose gel (Boehringer Mannheim, East Sussex, UK) and the reamplified target bands extracted from the gel as above. The eluted DNA was immediately ligated into a TOPO TA Cloning® vector (Invitrogen, Carlsbad, USA) before transformation in *Escherichia coli* TOP10F' One Shot™ cells (Invitrogen).

2.5. Colony screening

2.5.1. Stage I

Eight transformed (white) colonies from each band were grown up for 6 h in 200 µl LB broth containing ampicillin (Sigma, 50 mg/ml). One microlitre of this was subjected to PCR using the same conditions and nested primers as described above. One tenth (2 µl) of the completed PCR reaction was electrophoresed on a 2% standard agarose gel and one tenth on a 2% standard agarose gel containing HA red (Hanse Analytik GmbH, Bremen, Germany, 1 U/ml) to discern the differentially cloned products. The remainder of the PCR reaction was used to prepare duplicate dotblots on Hybond N+ membranes (Amersham, Little Chalfont, UK).

2.5.2. Stage II

The duplicate dotblots were probed with (a) the final differential display reaction and (b) the 'reverse-subtracted' differential display reaction. To make the 'reverse-subtracted' probe, the subtractive hybridisation step of the differential display

RT-PCR procedure was carried out using the original tester (treated) mRNA as the driver and the original driver (control) mRNA as the tester. Probing and visualisation were carried out using the ECL direct nucleic acid labelling and detection system (Amersham, Little Chalfont, UK) according to the manufacturer's instructions. Those clones that were positive for (a) but negative for (b), or showed a substantially larger positive signal with (a) compared to (b), were selected for DNA sequence analysis.

2.6. DNA sequencing

The remainder of the cultures (prepared in stage 1 screening) containing different cloning products (as discerned in the two screening steps) were grown up overnight in 5 ml LB broth containing ampicillin (50 mg/ml). A plasmid miniprep was prepared from each (Wizard Plus SV Minipreps DNA purification system, Promega) according to the manufacturer's instructions. The cloned inserts were sequenced on an automated ABI DNA sequencer (Applied Biosystems, Warrington, UK) using the M13 forward primer (GTAAAACGACGGCCAGT) or M13 reverse primer (AACAGCTATGACCATG).

2.7. Identification of differentially regulated genes

Gene sequences thus obtained were identified using the FASTA 3.0 programme (Lipman and Pearson, 1985; Pearson and Lipman, 1988) (<http://www.ddbj.nig.ac.jp/E-mail/homology.html>) to search all EMBL databases for matching DNA sequences. Each clone sequence was submitted in the forward and reverse direction, and the one returning the highest statistical probability of match to a known sequence was noted. Sequence homologies between our submitted clone sequence and the queried database sequence were determined (by FASTA) over a region of at least 60 base pairs.

2.8. RT-PCR analysis of selected candidate genes

cDNA sequences of the target genes were obtained from the NIH gene database (GenBank at

<http://www.ncbi.nlm.nih.gov/Web/Search/index.html>) and the computer programme GENE JOCKEY (BioSoft, Cambridge, UK) used to select primer pairs from these sequences. Where guinea pig sequences were available, rat and guinea pig sequences were aligned and primers chosen from regions of homology. If guinea pig sequences were not available, rat and human sequences were used. In cases where exact homology could not be found, the sequence from the rat was used. In the case of CD81 only, no rat or guinea pig sequences were available and so mouse and human sequences were aligned and a primer pair chosen from a region of homology. Primers (obtained from Gibco-BRL, Paisley, UK) were dissolved at a concentration of 50 pmol/ μ l in sterile distilled water and stored at -20°C . The primer pairs used plus other reaction parameters are shown in Table 1. mRNA was extracted (as described above) from all four treated animals and from three animals in the control group. Integrity of the eluted mRNA was confirmed on a 2% agarose gel, and the concentration and purity were measured using a Genequant II spectrophotometer (LKB, Bromma, Sweden) and then diluted to 10 ng/ μ l. One microlitre of this latter solution was used per RT-PCR reaction.

RT-PCR was carried out in a single tube (50 μ l) reaction using the Access RT-PCR system (Promega) according to manufacturer's instructions. In the kinetic and quantitative analyses, omission of RNA was used as a control for the presence of any contaminating DNA. After obtaining a PCR signal of the correct size and optimising the reaction conditions, each PCR product was digested with between two and four separate restriction enzymes. Specific restriction patterns were thus obtained, which further confirmed the identity of the PCR products as being the original target genes. Kinetic analysis (14–32 cycles) was then performed in each case to determine the location of the mid-log phase.

For the semi-quantitative analysis of each target gene, RT-PCR reactions were carried out in triplicate for each sample to reduce the effect of intertube RT-reaction variations (Kolls et al., 1993) and pipetting errors. For each gene, a mastermix containing enough reagents for three times

Table 1
Primer sequences and reaction conditions used in semi-quantitative RT-PCR analysis of selected genes

Transcript	Genbank accession No.	Primer sequences		Size of rat PCR product (bp)	Annealing temperature (°C) rat/guinea pig	No. of PCR cycles rat/guinea pig
		upstream	downstream			
Albumin	J00698 (rat)	TGGAGAGA-GAGC- CTTCAAGC	CTTAG- CAAGTCTCAGCAG CA	436	60/59	15/22
Bifunctional enzyme	K03249 (rat)	GCACC- CACTTCTTCT- CACCAGC	TGGCAATGATG- GTCCAGTAAGG	347	57/-	21/-
CYP2C11	J02657 (rat)	CCATCATGACC- CTGAGG	GAAGTCCCGAG- GATTGT	410	50/-	20/-
CYP4A1	M14972 (rat)	GATGGCTGCAC- CATGAG	GGCCTTTG- GATCTGATC	357	57/-	22/-
Catalase	M11670 (rat)	ACCAAATACTC- CAAGGCCAAAGG	GCCCTG- GTCAGTCTTG- TAATGG	450	63/-	27/-
CD81 (TAPA-1)	X59047 (mouse)	ATTTCGTCTTCTG	GCCTGGTCATA- GAACTGCTTCA	337	57/59	23/22
Contrapsin-like protease inhibitor	RNCCP23 (rat)	GCTGGCTGG GACTATGTGAG- CAATCAGAC	GTCTCTGGTTG- CAAGCT	341	50/-	20/-
Parathymosin- α (Zn ²⁺ binding protein)	X64053 (rat)	CGGCACCAT- GTCGGAGAAAGA	TTGTGTGTTCT- GCCCCACC	382	62/-	24/-
Transferrin	D38380 (rat)	AGCTGTGT- CAACTGT- GTCCAGG	GAGGAGAGGCC- GAACAGTTG- GAA	360	57/59	22/22
UDP-GT	U06273 (rat)	GGAT- GTCTGGGAAGTG	GCAGTTCAGC- TATCAGCT	495	50/-	23/-
DownUnknown-1	n/a	CGACGTTTC- CAAGGCA	TGTTGCGGCA- GAGTGA	318	55/-	25/-
Zn α 2glycoprotein	D21058 (rat)	CAAATAACA- GAAGCAGTG- GAGC	GACTTCCAC- CTCCATCCAGG	433	57/-	23/-

the number of samples (seven for rat, six for guinea pig) was prepared except that mRNA was omitted, the latter being added after aliquoting 49 μ l of the mastermix into an appropriate number of tubes. Amplification of albumin (the reference gene) was carried out in separate tubes since the mid-log phase of this gene is at a much lower cycle number than the target genes due to its high abundance. All RT-PCR products were analysed on 2% agarose gels containing 0.5 μ g/ml ethidium bromide. The target gene samples were loaded on the gel first and run in at 3 V/cm for 10 min. The corresponding albumin samples were then loaded and the gel run for a further 1/2 h. In this way, all

RT-PCR products from each target gene and albumin from the corresponding samples could be run on the same gel. Gels were photographed using type 665 posi-neg film (Sigma) and quantitation of the band intensity was carried out using a dual wavelength flying spot laser scanner densitometer (Shimadzu).

2.9. Statistical analysis

Statistical analysis of unpaired samples was carried out using the two-tailed Student's *t*-test. Values were considered statistically significant at $P < 0.05$ or less.

3. Results

3.1. Cloning and screening of transcripts

For both the rat and guinea pig experimental groups, cDNA subtraction was carried out in the forward (control driving tester) and reverse (tester driving control) directions to isolate both upregulated and downregulated mRNA species respectively. Using a standard primary hybridisation time of 8 h we obtained a substantial amount of non-specific products in all the final differential displays (data not shown). This background smearing was almost completely removed by reducing the primary hybridisation time to 4 h (CLONTECHniques, 1996). Fig. 1 shows the ddRT-PCR patterns of genes showing altered expression in rat and guinea pig liver following 3-day treatment with Wy-14,643. The profiles are unique for each species, and in each case the profile for the upregulated genes (control mRNA driving tester mRNA) is different to that obtained for the downregulated genes (tester mRNA driving control mRNA).

The practical outcome of the SSH method is that a series of differentially expressed genes is observed as a ladder on an agarose gel. The majority of these gene fragments fall within the 150–2000 bp range, with bands up to 5 Kbp occasionally being observed. Each band may theoretically consist of one or more products of similar size, as the gel has a maximum resolution

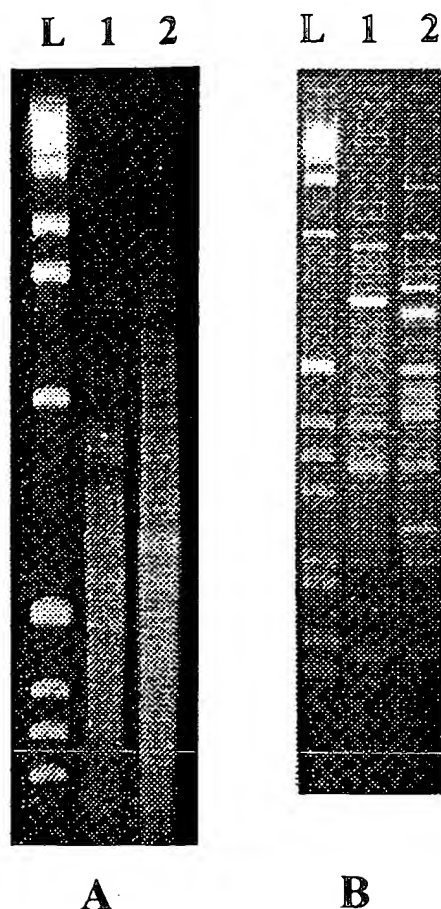


Fig. 1. Final displays of differentially expressed genes that were (1) upregulated and (2) downregulated in rat (A) and guinea pig (B) livers following 3-day treatment with Wy-14,643. mRNA extracted from control and treated livers was used to generate the differential displays using the PCR-Select cDNA subtraction kit (Clontech). Lane (L) is a 1 Kb DNA Ladder standard and 10 μ l of secondary PCR reaction were loaded in all other lanes.

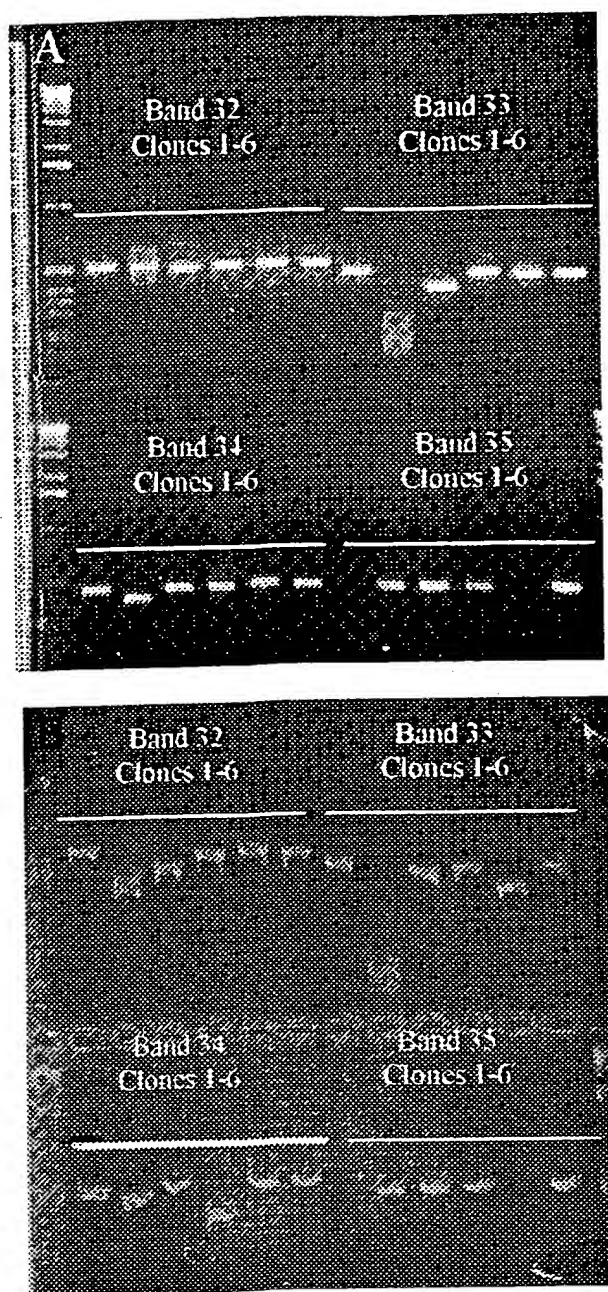


Fig. 2. Discrimination of different ddRT-PCR products having the same molecular size using HA-red. Gel (A) is a 2% standard agarose gel. Gel (B) is a 2% standard agarose gel containing 1 U/ml HA-red. Band numbers refer to the sequential bands (largest to smallest) extracted from the original display of genes upregulated in rat liver following 3-day treatment with Wy-14,643. Ten microlitres of each PCR reaction were loaded per lane.

of approximately 1.5% (3 bp per 200). In addition, there may be two or more products that are the same size, but have a different sequence.

Therefore some form of discrimination must be employed to isolate as many of these products as possible. HA-red screening (Geisinger et al., 1997) of a number of clones derived from each band provided a means to discriminate between different gene species of the same size. A typical example of such a gel is shown in Fig. 2. In total, 88 and 48 apparently different clones were obtained from the final differential expression patterns of upregulated and downregulated rat genes, respectively. Sixty nine and 89 apparently different clones were obtained from the final differential expression patterns of the upregulated and downregulated guinea pig genes, respectively.

Having identified as many different candidate gene products as possible in the screening step I, a second screening step was carried out on every clone to confirm those that represented true differentially expressed genes. This is necessary since no subtraction technique is 100% efficient. The approach we used, termed PCR-select differential screening (as recommended in Clontech's PCR-select cDNA subtraction kit protocol), utilises the forward and reverse subtractions as an aid to screening for the true differentially expressed genes (CLONTECHniques, 1997). Because these probes have already undergone subtraction, they have been enriched for differentially expressed genes and are therefore more sensitive than unsubtracted driver/tester cDNA probes for detecting true differential expression. All the clones that were isolated from each display were dotblotted and probed with the display from which they was obtained, plus the corresponding reverse-subtracted display. An example of such a blot is shown in Fig. 3. Clones corresponding to authentic differentially expressed mRNAs hybridised with the subtracted cDNA probe, but not the reverse-subtracted probe. We also included in the authentic positives, those clones that gave a substantially greater signal with the subtracted probe compared to the reverse-subtracted probe. False positives hybridised with either both probes or with neither probe. Of the original 88 upregulated and 48 downregulated rat clones selected for this screening step, 28 (32%) and 15 (31%) respectively, were found to be true positives. In the rat,

8 (100%) of the true positive upregulated genes (Table 2) and 11 (73%) of the true positive down-regulated genes (Table 3) were non-redundant. Of the original 69 upregulated and 89 downregulated guinea pig clones selected for this screening step, 8 (70%) and 37 (42%) respectively, were found to be true positives. Thirty six (75%) of the upregulated genes (Table 4) and 33 (89%) of the down-regulated genes (Table 5) were non-redundant.

3.2. Identification of clones

On sequence analysis it was found that some clones were unsequencable in the first instance (M13 forward primer) due to long polyA runs that appeared to prematurely terminate the sequencing reaction. These clones were therefore resequenced from the opposite direction using the M13 reverse primer. Those xenobiotic-modulated gene products identified to date are listed in Tables 2 and 3 (rat) and Tables 4 and 5 (guinea pig).

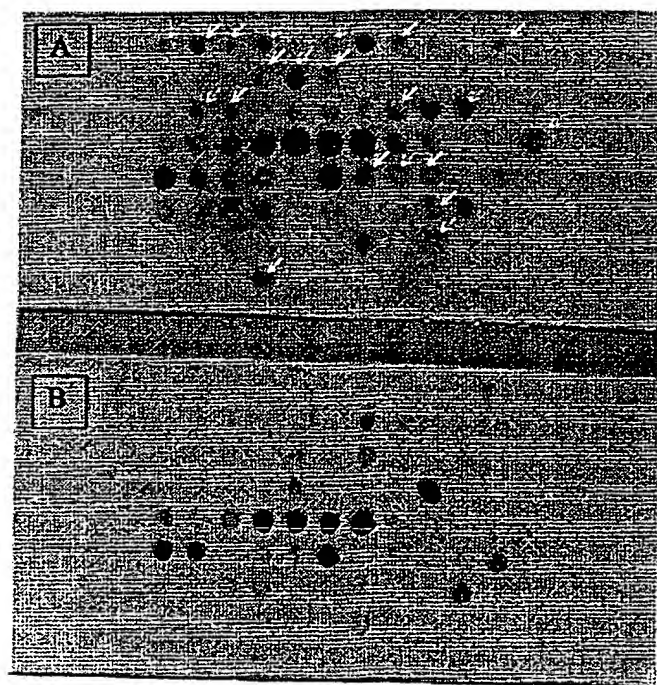


Fig. 3. Dot blots of clones of putative upregulated gene species isolated from guinea pig liver following 3-day treatment with Wy-14,643. All clones identified in the stage I screening step (see methods) were blotted and probed with (A) the differential display from which they originated (control driving treated) and (B) the reverse subtraction (treated driving control). Arrows indicate some of the true differentially expressed clones.

Table 2

Identification of genes that were upregulated in male rat liver following 3-day treatment with WY-14,643

FASTA-EMBL gene identification (rat unless otherwise stated)	Accession No.	Sequence homology ^a (%)
Carnitine octanoyl transferase	RN26033	99
NCI_CGAP_Li1 (<i>H. sapiens</i>) (EST ^b)	HS1275949	98
Peroxisomal enoyl hydratase-like protein	RN08976	98
Liver fatty acid binding protein	V01235	96
Soares mouse p3NMF19.5 <i>M. musculus</i> cDNA clone	AA038051	96
Cytochrome p450IVA1	RNCYPLA	94
Mit. 3-hydroxyl-3-methylglutaryl CoA synthase	RNHMGCOA	94
Rabgeranylgeranyl transferase component B	RNRABGERA	94
Genes for 18S, 5.8S, and 28S ribosomal RNAs	RNRRNA	94
Carnitine acetyl transferase (mouse)	MMRNACAR	92
Soares mouse NML (EST)	MM1157113	92
Bone marrow stromal fibroblast (<i>H. sapiens</i>) cDNA clone HBMSF2E4 (EST)	AA545726	92
7.5dpc embryo (mouse) (EST)	AA408192	92
Alpha-1-macroglobulin	RNALPH1M	91
Transferrin	RNTRANSA	91
Lecithin:cholesterol acyltransferase	RNU62803	90
Zn-α2-glycoprotein	RNZA2GA	90
Serum albumin	RNJALBM	89
Fructose-1,6-bisphosphate 1-phosphohydrolase	RNFBP	88
Soares mouse melanoma (EST) (S ^c)	AA124706	88
Soares mouse 3NbMS (EST) (AS ^c)	AA154039	88

Table 2 (Continued)

FASTA-EMBL gene identification (rat unless otherwise stated)	Accession No.	Sequence homology ^a (%)
17- β -hydroxysteroid dehydrogenase	RN17BHDT2	87
Soares mouse p3NMF19.5 (EST)	AA038051	87
Peroxisomal enoyl-CoA:hydratase-3-hydroxyacyl CoA bifunctional enzyme	RNPECOA	85
Integral membrane protein, TAPA-1 (CD81) (mouse)	S45012	81
Soares mouse lymph node (EST)	MMAA88445	81
<i>H. sapiens</i> (clone zap128) mRNA	L40401	76
Lysophospholipase homologue (human)	HSU67963	76
Soares mouse lymph node (EST)	AA217044	74

^a Refers to the nucleotide sequence homology between the cloned band isolated from the differential display and the corresponding gene derived from the EMBL gene sequence bank.

^b EST is 'expressed sequence tag' — a gene of as yet unknown identity and function.

^c Where sequence homologies were equal in both directions of the isolated band, both the sense (S) and antisense (A) identities are given.

Table 3

Identification of genes that were downregulated in male rat liver following 3-day treatment with Wy-14,643

FAST-EMBL gene identification (rat unless otherwise stated)	Accession No.	Sequence homology ^a (%)
NCI_CGAP_Li1 (<i>H. sapiens</i>) (EST ^b)(S ^c)	AA484528	99
NCI_CGAP_Pr1 (<i>H. sapiens</i>) (EST)(AS ^c)	AA469320	99
UDP-glucuronosyl-transferase (UGT2B12)	RN06273	98
Complement component c3	RNC3	96
Soares mouse placenta (S)	AA023305	96
Ape (chimpanzee) 28S rRNA (AS)	PTRGMC	96
Rat CYP2C11	RNCYPM1	95
Ribosomal protein S5	RNRPS5	94
Transthyretin	RNTTHY	94
Contrapsin-like protease inhibitor	RNCCP23	89
Prostaglandin F2a (S)	RN26663	84
β -2-microglobulin (AS)	RNB2MR	84
Apolipoprotein C-III	RNAPOA02	82
Parathymosin-alpha (zinc2 ⁺ -binding protein)	RN11ZNBP	75

^a Refers to the nucleotide sequence homology between the cloned band isolated from the differential display and the corresponding gene derived from the EMBL gene sequence bank.

^b EST is 'expressed sequence tag' — a gene of as yet unknown identity and function.

^c Where sequence homologies were equal in both directions, both the sense (S) and antisense (A) identities are given.

In all cases, both the forward and reverse sequence of the target clones were analysed and the gene having the highest statistical homology noted.

3.3. RT-PCR analysis of selected clones

The results of a typical RT-PCR semi-quantitation experiment for transferrin in the rat is given in Fig. 4 and the results for a total of 12 selected genes in both the rat and guinea pig are shown in Table 6.

4. Discussion

It is now apparent that all cancers arise from accumulated genetic changes within the cell. Although documenting and explaining these changes presents a formidable obstacle to understanding the different mechanisms of carcinogenesis, the experimental methodology is now available to begin attempting this difficult challenge. In order to begin the elucidation of the molecular mechanisms involved in non-genotoxic hepatocarcino-

genesis, we have used SSH to identify a number of genes that are upregulated or downregulated in male rat and guinea pig livers following short term exposure to the PP, Wy-14,643. We have used the rat model to represent a species susceptible to the non-genotoxic carcinogenic effect of PPs and the guinea pig as a resistant species (Orton et al., 1984; Rodricks and Turnbull, 1987;

Lake et al., 1989; Makowska et al., 1992; Lake et al., 1993).

Gurskaya et al. (1996), who originally developed the SSH technique, cloned the products of the secondary PCR reaction and screened a small number of randomly selected colonies for differentially expressed clones using northern hybridisation. However, we decided against this approach

Table 4
Identification of genes that were upregulated in male guinea pig liver following 3-day treatment with WY-14,643

FASTA-EMBL gene identification (guinea pig unless otherwise stated)	Accession No.	Sequence homology ^a (%)
Carboxylesterase	AB010634	97
Complement C3 protein (GPC3)	M34054	97
Cytosolic aldehyde dehydrogenase (sheep)	U12761	92
Catalase (human)	X04076	89
Mitochondrial aspartate aminotransferase (pig)	M11732	89
Elongation factor-1-alpha (rabbit)	X62245	88
NCI_CGAP_Br2 <i>H. sapiens</i> cDNA clone (EST) (Similar to chick mit. phosphoenolpyruvate carboxykinase)	AA587436	87
Alpha-1-antiproteinase S	M57270	83
10-formyltetrahydrofolate dehydrogenase (rat)	M59861	83
Ribosomal protein L6 (rat)	X87107	83
Moares pregnant uterus Nb (EST) (mouse)	AA156847	83
Mitochondrial citrate transport protein (human)	L77567	80
Cytoplasmic chaperonin hTRiC5 (human)	U17104	80
Alpha-1-antiproteinase F	M57271	77
Heterogeneous nuclear ribonuclearprotein c1/c2 (human)	D28382	77
Moares parathyroid tumour (EST) (similar to human serum albumin precursor)	AA860651	76
Stratagene mouse kidney (EST)	AA107327	75
Moares parathyroid tumour NbHPA human cDNA (EST)	AA860653	74
Moares mouse mammary gland (EST)	AA619297	74
cDNA clone 15 004 (EST) (human)	H01826	74
Moares senescent fibroblasts (EST) (mouse)	W52190	74
Preproalbumin (human)	E04315	72
cDNA clone 73 169 (EST) (human)	T56624	72
Vitamin D-binding protein (human)	L10641	71
ApoH gene (exon 8) (human)	Y11498	71
MSRL flow sorted chromosome	B05457	71
Moares foetal liver spleen (EST) (mouse)	AA009524	71
Moares foetal heart NbMH19W (EST) (mouse)	AA009421	69
Moares foetal heart NbHH19W <i>H. sapiens</i> cDNA clone (EST)	W94377	67
Phenylalanine hydroxylase (human)	U49897	67
Proline-5-carboxylate dehydrogenase (human)	U24266	66
Glutathione-S-transferase homologue (human)	U90313	65
NCI_CGAP_GCBI (EST) (human)	AA769294	65
Protective protein (human)	M22960	64
Clone 27 375 (EST) (human)	N37046	62
Stratagene colon (# 937 204) <i>H. sapiens</i> cDNA clone (EST)	AA149777	62

^a Refers to the nucleotide sequence homology between the cloned band isolated from the differential display and the corresponding gene derived from the EMBL gene sequence bank.

Table 5
Identification of genes that were downregulated in male guinea pig liver following 3-day treatment with WY-14,643

FASTA-EMBL gene identification (guinea pig unless otherwise stated)	Accession No.	Sequence homology ^a (%)
Complement C3 protein	M34054	97
Murinoglobulin	D84339	95
Alpha-1-antitrypsin	M57271	88
Elongation factor-1 (rabbit)	X62245	89
Coupling protein G (human)	X04409	88
NCI_CGAP_Ov1 (EST ^b) (human)	AA586309	87
Lecithin:cholesterol acetyl transferase (rabbit)	D13668	85
Aldolase B (human)	X00270	84
Anti-thrombin III (human)	E00116	80
Phenylalanine hydroxylase (human)	K03020	80
Inter- α -trypsin inhibitor (human)	D38595	79
Normalised rat muscle (EST) (S ^c)	AA849753	78
Normalised rat ovary (EST) (AS ^c)	AA801059	78
Complement factor Ba fragment (human)	X00284	77
Dihydrodiol dehydrogenase (human)	U05598	76
Spot14 gene (thyroid-inducible hepatic protein)(human)	Y08409	75
BAC clone 174p12 (human)	AC004236	75
Mitochondrial aldehyde dehydrogenase (human)	X05409	74
Preproalbumin (human)	E04315	74
NCI_CGAP_Pr9 (EST) (human) (S)	AA533142	74
Normalised rat placenta (EST) (AS)	AA851197	74
Heparin sulfate proteoglycan (human)	J04621	73
cDNA clone 33 992 (EST) (human)	R24330	73

Table 5 (Continued)

FASTA-EMBL gene identification (guinea pig unless otherwise stated)	Accession No.	Sequence homology ^a (%)
Retinol dehydrogenase (rat)	U33501	71
TAPA-1 integral membrane protein (CD81) (mouse)	S45012	71
Complement component c5s	M35525	70
Apolipoprotein B (pig)	L11235	69
cDNA clone 143 918 (EST) (human)	R76742	68
α -fibrinogen (human)	K02569	68
Soares foetal liver spleen INF (mouse)	W03726	68
Barstead bowel (EST) (mouse)	AA232049	67
UDP glucuronosyl transferase (cat)	AF0309137	66
Myeloid leukaemia cell differentiation protein (MCL-1) (human) (S)	L08246	65
STS SHGC-34 987 (human) (AS)	hu-G27984	65
Soares mouse 3NME125	AA222798	64
Stratagene mouse embryonic (EST) (S)	AA199420	64
Rad 52 (mouse)	AF004854	63

^a Refers to the nucleotide sequence homology between the cloned band isolated from the differential display and the corresponding gene derived from the EMBL gene sequence bank.

^b EST is 'expressed sequence tag' — a gene of as yet unknown identity and function

^c Where sequence homologies were equal in both directions, both the sense (S) and antisense (A) identities are given.

for several reasons: (1) the kinetics of ligation and transformation favour the isolation of smaller PCR products, thereby producing a misrepresentation of larger gene products; (2) northern blot analysis is notoriously insensitive and is unlikely to confirm expression of rare transcripts; (3) there is no measurable end point to the screening of clones produced in this way other than to analyse every transformed colony. We used instead an alternative approach; after running out the differ-

ential display on a high-resolution agarose gel (Fig. 1) and overstaining with SYBR Green I to enhance visualisation, the composite bands were individually extracted, reamplified and cloned. However, it has been well documented that single bands from differential displays often contain a heterogeneous mixture of different products (Mathieu-Daude et al., 1996; Smith et al., 1997). This is because polyacrylamide gels cannot discriminate between DNA sequences that differ in size by less than about 0.2% (Sambrook et al., 1989). High-resolution agarose gels such as those used in this work are even less sensitive, normally only discriminating products that differ in size by at least 1.5%. The use of the HA-red screening step enables resolution of identical or nearly identical sequences based on their AT content (Wawer et al., 1995) and is sensitive down to < 1% difference. Furthermore, it is rapid, technically simple and does not require the use of radiolabels. Geisinger et al. (1997) originally demonstrated the usefulness of using HA-red to identify different products cloned from the same band of an RNA differential display experiment by simultaneously running them in normal agarose (to discriminate by size) and in normal agarose containing HA-red (to discriminate by AT content). We have found that this approach is equally useful for identifying different gene species cloned from the same band of our SSH display.

Diatchenko et al. (1996) reported that SSH is highly efficient at producing differentially expressed gene species. However, we also included a second screening step to further confirm that the clones isolated from the differential display were indeed differentially expressed. Duplicate dotblots of the candidate clones were blotted with the display from which they were originally isolated and with the 'reverse subtraction' display. To make the reverse-subtracted probe, the subtractive hybridisation step of the procedure was carried out using the original tester cDNA as a driver, and the original driver cDNA as a tester. In this way, clones that are false positives can be identified through their presence in both blots. Such false positives most commonly arise through having a very high abundance in the initial sample or unusual hybridisation properties (Li et al., 1994).

Although the SSH method itself has been shown to be efficient, and despite the screening step that we included, there is an important caveat to bear in mind — namely that it is important that all clones be considered only as 'candidates' until the actual abundance of their mRNA is quantitated in treated and control samples. Towards this end, we examined the expression of a limited number of clones using semi-quantitative RT-PCR. Albumin was used as the reference gene as we have previously found that the expression of this gene does not appear to change with the treatment regime that we used (Fig. 4, and data not shown). There are a number of interesting points to note from our results. The first is the presence of genes that serve as appropriate positive controls in the upregulated and downregulated series. For example, in the rat it can be seen that CYP4A1 expression increases 14-fold following treatment. Although CYP4A1 mRNA expression levels following WY-14,643 treatment have not been previously reported in this model, the figure compares favourably with that recorded by Bell et al. (1991), who used RNase-protection to quantitate CYP4A1 in rat liver following treatment with methylofenapate, another PP. In addition, we also confirmed that the peroxisomal enoyl-CoA:hydratase-3-hydroxyacyl-CoA bifunctional enzyme is also upregulated 9-fold, in agreement with the findings of Chen and Crane (1992).

A number of genes were downregulated following Wy-14,643 exposure, including CYP2C11 expression. Corton et al. (1997) reported similar findings and suggested that this may in part explain why male rats exposed to Wy-14,643 and some other PPs have high serum estradiol levels, as estradiol is a substrate for CYP2C11. We have also shown that the expression of contrapsin-like protease inhibitor (CLPI) was downregulated by Wy-14,643. This has not previously been reported, and we suggest that it may be linked to a requirement for increased availability of amino acids to accommodate the hepatomegaly induced by treatment. Although little is known of the function of parathymosin- α , (zinc²⁺-binding protein) it has been shown to interact with the globular domain of histone H1, suggesting a role in histone function (Kondili et al., 1996). In contrast to the

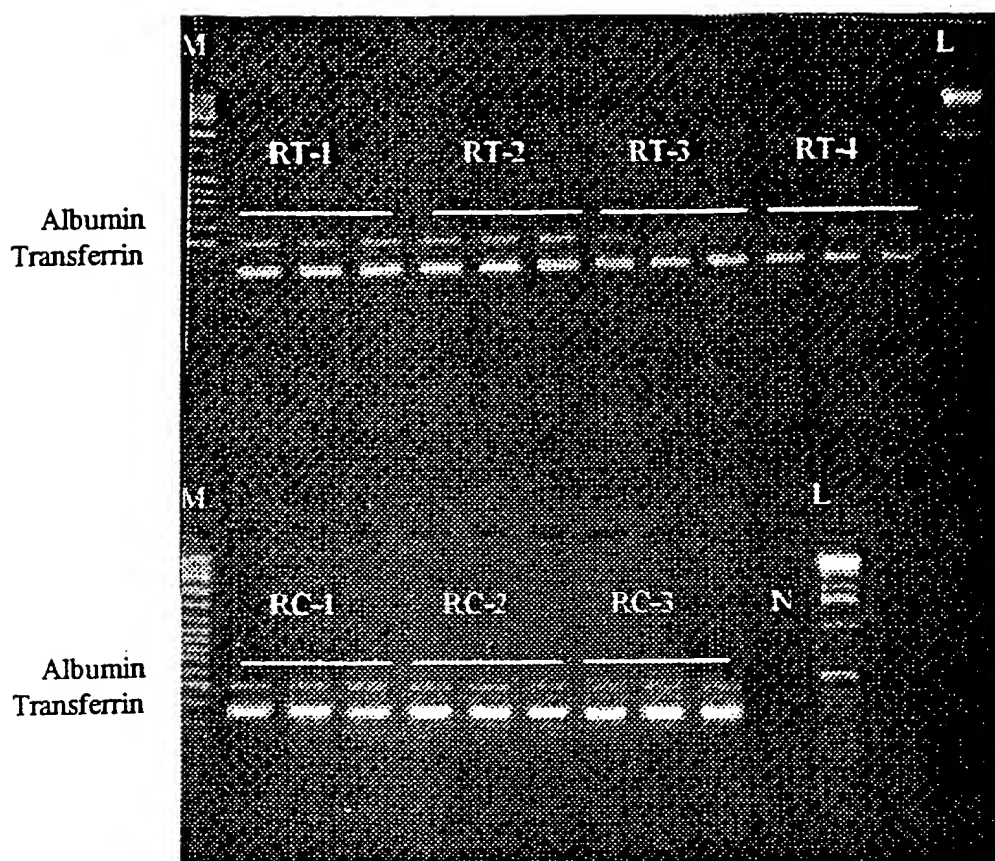


Fig. 4. Semi-quantitative RT-PCR experiment showing relative decrease in expression of transferrin in treated rat liver (RT-1 to RT-4) compared to controls (RC-1 to RC-3). An equal amount of mRNA was used in each reaction (10 ng), and each sample was quantitated in triplicate to reduce the effects of inter-tube variation. N is negative control (no mRNA). Lane M is a 100 bp ladder and lane L is a 1 Kb DNA ladder.

downregulation observed in this work, other studies have shown that parathymosin- α expression is elevated in breast cancer (Tsitsilonis et al., 1993, 1998), with the implication that parathymosin- α may somehow be involved in regulating cell proliferation by more than one mechanism. Transferrin has previously been shown to be downregulated in rat liver by hypolipidemic PPs (Hertz et al., 1996). It is therefore interesting to note that we isolated a clone identified as transferrin from the upregulated display profile. Since we confirmed by RT-PCR that transferrin is in fact downregulated in the rat (Fig. 4), we conclude that transferrin was either a false positive or was incorrectly identified. It could also be that we have isolated a close relative, splice variant or isoform of transferrin, which demonstrates a different expression profile under these experimental conditions. Further investigations are therefore

required to determine which of these possibilities are correct.

One of our most intriguing observations was that one gene, CD81, appeared to be upregulated in rat liver but downregulated in guinea pig liver following Wy-14,643 exposure. CD81 is a widely expressed cell surface protein that is involved in a large number of cellular functions, including adhesion, activation, proliferation and differentiation (reviewed by Levy et al., 1998). Since all of these functions are altered to some extent in carcinogenesis, it is perhaps an important observation that CD81 expression is differentially regulated in a resistant and sensitive species exposed to a non-genotoxic carcinogen.

Albumin and ribosomal genes appear common to all differential displays and are thus undesirable false positives. However, due to their high expression in the liver, they are difficult to re-

move. We also noted a number of gene species, particularly in the guinea pig, which were common to both upregulated and downregulated profiles. Again, the most likely reason for these having arisen is their high abundance.

A relatively large number of upregulated and downregulated genes were isolated from guinea pig liver following Wy-14,643 exposure. However, the guinea pig genome has been relatively poorly characterised and so many of the clones were identified as resembling genes or ESTs from other species. Without full-length sequence data it is difficult to ascertain the accuracy of the assigned identities and this must be borne in mind when utilising data such as this, for example, in designing effective primers for RT-PCR studies. Although the actual isolated clone sequences can be used to do this, their relatively small size often restricts the ability to design effective primers. In addition, as we observed with transferrin, using a published full-length sequence may help to identify false positives.

By comparing the expression profiles of genes showing altered expression in a PP-sensitive species (rat) with a PP-resistant species (guinea pig), it was our aim to identify genes that are mechanistically relevant to the non-genotoxic hepatocarcinogenic action of Wy-14,643. However, few of the genes that we have isolated were common to both the rat and the guinea pig. This suggests either that the molecular mechanisms of response in these two species are so different that few genes are commonly regulated in response to Wy-14,643 exposure, or that we have recovered only a small proportion of those genes that have altered expression. The latter seems the more likely scenario since it is perceived that one of the main problems of subtractive hybridisation and other differential expression technologies is the inability to consistently isolate rare gene transcripts (Bertioli et al., 1995). This is potentially problematic in that weakly expressed genes may play an important role in regulating key cellular processes, and that the majority of mRNA species are classified as

Table 6
Semi-quantitative RT-PCR analysis of selected gene species in the rat and guinea pig^a

Transcript	Putative change of expression following treatment according to dotblot		Change according to RT-PCR quantitation	
	Rat	Guinea pig	Rat	Guinea pig
Albumin	N/A	N/A	No change	No change
Bifunctional enzyme	Up	N/A	Upregulated* (9 ×)	N/O
CYP2C11	Down	N/A	Downregulated* (Abolished)	N/D
CYP4A1	Up	N/A	Upregulated* (14 ×)	N/D
Catalase	N/A	Up	No change	N/O
CD81 (TAPA-1)	Up	Down	N/O	Upregulated** (1.4 ×)
Contrapsin-like protease inhibitor	Down	N/A	Downregulated** (0.5 ×)	N/D
Parathymosin-α (zinc ²⁺ binding protein)	Down	N/A	Downregulated** (0.6 ×)	N/D
Transferrin	Up	N/A	Downregulated* (0.5 ×)	No change
UDP-Glucuronosyl transferase	Down	N/A	Downregulated** (0.2 ×)	N/O
DownUnknown-1	Down	N/A	No change (<i>P</i> = 0.06)	N/D
Zn-α2-glycoprotein	Up	N/A	No change	N/O

^a N/A, not applicable; N/O, not optimised; N/D, not done.

* *P* < 0.0005;

** *P* < 0.05.

'rare' in abundance (Bertioli et al., 1995). However, in their original paper describing the SSH technique, Gurskaya et al. (1996) demonstrated that SSH can enrich rare molecules between 1000- and 5000-fold in a single round of hybridisation. Unfortunately, due to high background smearing in our initial experiments (which hindered identification of single bands), we were compelled to reduce the primary hybridisation time to only 4 h — a step that theoretically is likely to reduce the number of rare sequences (CLONTECHniques, 1996). Furthermore, it has been claimed by the manufacturers that, whilst this technique can identify changes as small as 1.5-fold between the driver and tester populations, it is best suited to the isolation of genes that show a greater than 5-fold increase (CLONTECHniques, 1996). In addition, where tester and driver contain genes with large and small differences in abundance, the SSH method will be biased towards identifying those genes with the large differences (CLONTECHniques, 1996). Thus, it is most probable that we have not isolated all of the more rarely expressed transcripts and those demonstrating small changes in expression.

One problem that remains is identifying the function of genes isolated in SSH experiments as described herein, some of which may be crucial to the process of carcinogenesis, and are, to date, unidentified. However, we have provided evidence herein that SSH can be used to begin the process of characterising the extent and importance of altered gene expression in response to a chemical stimulus. The developments of this approach should include characterisation of temporal and dose responses, and functional analysis studies including knockout mice. In combination, such studies should make a significant contribution to our understanding of the molecular mechanisms of action and physiological relevance of gene regulation in non-genotoxic hepatocarcinogenesis. It should then be possible to ascertain whether differentially expressed genes are causally or casually related to the chemical-induced toxicity, and therefore a substantial mechanistic advance.

It is clear that there are also broader applications for this experimental approach that go beyond understanding the molecular mechanisms of

peroxisome-proliferator induced non-genotoxic hepatocarcinogenesis in rodents. The potential medical and therapeutic benefits of elucidating the molecular changes that occur in any given cell in progressing from the normal to the carcinogenic (or other diseased, abnormal or developmental) state are very substantial. Notwithstanding the lack of complete functional identification of altered gene expression, such gene profiling studies described herein essentially provides a 'fingerprint' of each stage of carcinogenesis, and should help in the elucidation of specific and sensitive biomarkers for different types of cancer. Amongst other benefits, such fingerprints and biomarkers could help uncover differences in histologically identical cancers, and provide diagnostic tests for the earliest stages of neoplasia. In addition, the genes identified by this approach may be incorporated into gene-chip DNA-arrays, thus providing a standard genetic fingerprint for a particular toxin treatment in a particular species. Interrogation of these gene arrays for an unknown compound that has a similar pattern to the known reference chemical would then provide evidence that the unknown may have a toxicity profile similar to the 'standard' fingerprint, thereby serving as a mechanistically relevant platform for further detailed investigations.

Acknowledgements

This work was funded by Rhone-Poulenc Agrochemicals, Nice, France.

References

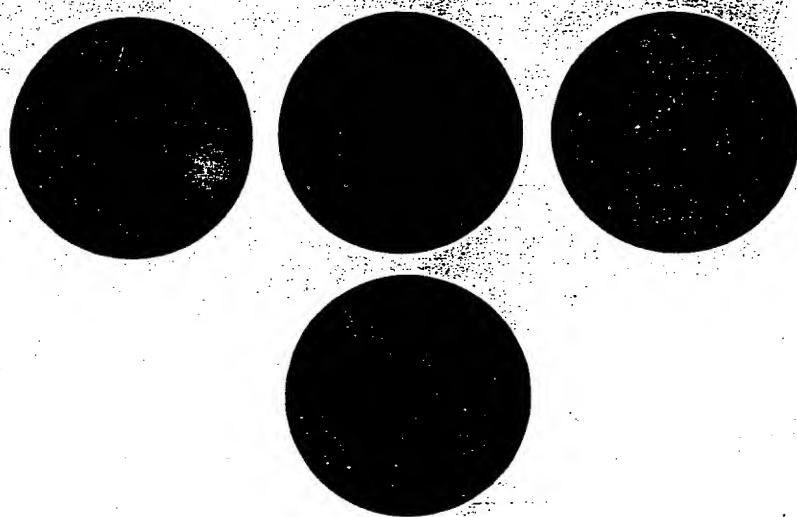
- Anderson, N.L., Esquer-Blasco, R., Richardson, F., Foxworthy, P., Eacho, P., 1996. The effects of peroxisome proliferators on protein abundances in mouse liver. *Toxicol. Appl. Pharmacol.* 137, 75–89.
- Ashby, J., 1992. Prediction of non-genotoxic carcinogenesis. *Toxicol. Lett.* 64–65, 605–612.
- Bell, D.R., Bars, R.G., Gibson, G.G., Elcombe, C.R., 1991. Localisation and differential induction of cytochrome P450IVA and acyl coA oxidase in rat liver. *Biochem. J.* 275, 247–252.
- Bertioli, D.J., Schlichter, U.H.A., Adams, M.J., Burrows, P.R., Steinbiss, H.-H., Antoniow, J.F., 1995. An analysis of

- differential display shows a strong bias towards high copy number mRNAs. *Nucleic Acid Res.* 23 (21), 4520–4523.
- Cattley, R.C., Kato, M., Popp, J.A., Teets, V.J., Voss, K.S., 1994. Initiator-specific promotion of hepatocarcinogenesis by Wy-14,643 and clofibrate. *Carcinogenesis* 15 (8), 1763–1766.
- Chen, N., Crane, D.I., 1992. Induction of the major integral membrane protein of mouse liver peroxisomes by peroxisome proliferators. *Biochem. J.* 283, 605–610.
- CLONTECHniques, 1996. Technical Tips: Clontech PCR-Select cDNA Subtraction, October 25; application notes.
- CLONTECHniques, 1997. PCR-Select Differential Screening Kit — The Next Step After Clontech PCR-Select cDNA Subtraction. XII(2), 18–19, application notes.
- Corton, J.C., Bocos, C., Moreno, E.S., Merritt, A., Cattley, R.C., Gustaffson, J.A., 1997. Peroxisome proliferators alter the expression of estrogen-metabolising enzymes. *Biochimie* 79, 151–162.
- Davis, M., Cohen, D.I., Nielson, E.A., Steinmetz, M., Paul, W.E., Hood, L., 1984. Cell-type-specific cDNA probes and the murine I region: the localisation and orientation of Ad/a. *Proc. Natl. Acad. Sci. USA* 81, 2194–2198.
- Diatchenko, L., Lau, Y.-F.C., Campbell, A.P., Chenchik, A., Moqadam, F., Huang, B., Lukyanov, K., Gurskaya, N., Sverdlov, E.D., Siebert, P.D., 1996. Suppression subtractive hybridisation: a method for generating differentially regulated or tissue-specific cDNA probes and libraries. *Proc. Natl. Acad. Sci. USA* 93, 6025–6030.
- Duguid, J., Dinauer, M., 1990. Library subtraction of in vitro cDNA libraries to identify differentially expressed genes in scrapie infection. *Nucleic Acid Res.* 18 (9), 2789–2792.
- Geisinger, A., Rodriguez, R., Romero, V., Wettstein, R., 1997. A simple method for screening cDNAs arising from the cloning of RNA differential display bands. Elsevier trends journals technical tips online, <http://tto.trends.com>, document number T01110
- Guimaraes, M.J., Lee, F., Zlotnik, A., McClanahan, T., 1995. Differential display by PCR: novel findings and applications. *Nucleic Acid Res.* 23 (10), 1832–1833.
- Gurskaya, N.G., Diatchenko, L., Chenchik, P.D., Siebert, P.D., Khaspekov, G.L., Lukyanov, K.A., Vagner, L.L., Ermolaeva, O.D., Lukyanov, S.A., Sverdlov, E.D., 1996. Equalising cDNA subtraction based on selective suppression of polymerase chain reaction: cloning of Jurkat cell transcripts induced by phytohemagglutinin and phorbol 12-myristate 13-acetate. *Anal. Biochem.* 240, 90–97.
- Hayashi, F., Tamura, H., Yamada, J., Kasai, H., Suga, T., 1994. Characteristics of the hepatocarcinogenesis caused by dehydroepiandrosterone, a peroxisome proliferator, in male F-344 rats. *Carcinogenesis* 15 (19), 2215–2219.
- Hedrick, S.M., Cohen, D.I., Nielsen, E.A., Davis, M.M., 1984. Isolation of cDNA clones encoding T cell-specific membrane-associated proteins. *Nature* 308 (8), 149–153.
- Hertz, R., Seckbach, M., Zakin, M.M., Bar-Tana, J., 1996. Transcriptional suppression of the transferrin gene by hypolipidemic peroxisome proliferators. *J. Biol. Chem.* 271 (1), 218–224.
- Hubank, M., Schatz, D.G., 1994. Identifying differences in mRNA expression by representational difference analysis. *Nucleic Acid Res.* 22 (25), 5640–5648.
- Human and Experimental Toxicology, 1994. *Hum. Exp. Toxicol.* 13 (Suppl. 2) (entire issue).
- Kolls, J., Dsininger, P., Cohen, C., Larson, J., 1993. cDNA equalisation for reverse transcription-polymerase chain reaction quantitation. *Anal. Biochem.* 208, 264–269.
- Kondili, K., Tsolas, O., Papamarcaki, T., 1996. Selective interaction between parathymosin and histone H1. *Eur. J. Biochem.* 242 (1), 67–74.
- Lake, B.G., Evans, J.G., Gray, T.J.B., Korosi, S.A., North, C.J., 1989. Comparative studies of nafenopin-induced hepatic peroxisome proliferation in the rat, Syrian hamster, guinea pig and marmoset. *Toxicol. Appl. Pharmacol.* 99, 148–160.
- Lake, B.G., Evans, J.G., Cunningham, M.E., Price, R.J., 1993. Comparison of the hepatic effects of Wy-14,643 on peroxisome proliferation and cell replication in the rat and Syrian hamster. *Environ. Health Perspect.* 101 (S5), 241–248.
- Levy, S., Todd, S.C., Maecker, H.T., 1998. CD81 (TAPA-1): a molecule involved in signal transduction and cell adhesion in the immune system. *Annu. Rev. Immunol.* 16, 89–109.
- Li, W.B., Gruber, C.E., Lin, J.J., D'Alessio, J.M., Jessee, J.A., 1994. The isolation of differentially expressed genes in fibroblast growth factor stimulated BC3H1 cells by subtractive hybridization. *BioTechniques* 16, 722–729.
- Liang, P., Pardee, A.B., 1992. Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science* 257 (5072), 967–971.
- Liang, P., Averboukh, L., Keyomarsi, K., Sager, R., Pardee, A.B., 1992. Differential display and cloning of messenger RNAs from human breast cancer versus mammary epithelium. *Cancer Res.* 52, 6966–6968.
- Lipman, D.J., Pearson, W.R., 1985. Rapid and sensitive protein similarity searches. *Science* 227, 1435–1441.
- Makowska, J.M., Gibson, G.G., Bonner, F.W., 1992. Species differences in ciprofibrate induction of hepatic cytochrome P450IVA1 and peroxisome proliferation. *J. Biochem. Toxicol.* 7, 183–191.
- Marsman, D.S., Cattley, R.C., Conway, J.G., Popp, J.A., 1988. Relationship of hepatic peroxisome proliferation and replicative DNA synthesis to the hepatocarcinogenicity of the peroxisome proliferators di-(2-ethylhexyl)phthalate and [4-chloro-6-(2,3-xylidino)-2-pyrimidinylthio]acetic acid (Wy-14,643) in rats. *Cancer Res.* 48, 6739–6744.
- Mathieu-Daude, F., Cheng, R., Welsh, J., McClelland, M., 1996. Screening of differentially amplified cDNA products from RNA arbitrarily primed PCR fingerprints using single strand conformation polymorphism (SSCP) gels. *Nucleic Acid Res.* 24 (8), 1504–1507.
- Orton, T.C., Adam, H.K., Bentley, M., Holloway, B., Tucker, M.J., 1984. Clobuzarit: species differences in the morphological and biochemical response of the liver following chronic administration. *Toxicol. Appl. Pharmacol.* 73, 138–151.

- Parodi, S., 1992. Non-genotoxic factors in the carcinogenic process: problems of detection and hazard evaluation. *Toxicol. Lett.* 64–65, 621–630.
- Pearson, W.R., Lipman, D.J., 1988. Imported tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* 85, 2444–2448.
- Rockett, J.C., Esdaile, D.J., Gibson, G.G., 1997. Molecular profiling of non-genotoxic hepatocarcinogenesis using differential display reverse transcription-polymerase chain reaction (ddRT-PCR). *Eur. J. Drug. Metab. Pharmacokinet* 22 (4), 329–333.
- Rodricks, J.V., Turnbull, D., 1987. Inter-species differences in peroxisomes and peroxisome proliferation. *Toxicol. Ind. Health* 3, 197–212.
- Sambrook, J., Fritsch, E.F., Maniatis, T., 1989. In: Ford, N., Nolan, C., Ferguson, M. (Eds.), *Molecular Cloning — A Laboratory Manual*, second ed. Cold Spring Harbor Laboratory Press, New York.
- Sargent, T., Dawid, I., 1983. Differential gene expression in the gastrula of *Xenopus laevis*. *Science* 222, 135–139.
- Smith, N.R., Li, A., Aldersley, M., High, A.S., Markham, A.F., Robinson, P.A., 1997. Rapid determination of the complexity of cDNA bands extracted from DDRT-PCR polyacrylamide gels. *Nucleic Acid Res.* 25 (17), 3552–3554.
- Tsitsilonis, O.E., Stiakakis, J., Koutselinis, A., Gogas, J., Markopoulos, C., Yialouris, P., Bekris, S., Panoussopoulos, D., Kiortsis, V., Voelter, W., Haritos, A.A., 1993. Expression of alpha-thymosins in human tissues in normal and abnormal growth. *Proc. Natl. Acad. Sci. USA* 90 (20), 9504–9507.
- Tsitsilonis, O.E., Bekris, E., Voutsas, I.F., Baxevanis, C.N., Markopoulos, C., Papadopoulou, S.A., Kontzoglou, K., Stoeva, S., Gogas, J., Voelter, W., Papamichail, M., 1998. The prognostic value of alpha-thymosins in breast cancer. *Anticancer Res.* 18 (3A), 1501–1508.
- Wan, J.S., Sharp, S.J., Poirier, G.M.-C., Wagaman, P.C., Chambers, J., Pyati, J., Hom, Y.-L., Galindo, J.E., Huvar, A., Peterson, P.A., Jackson, M.R., Erlander, M.G., 1996. Cloning differentially expressed mRNAs. *Nat. Biotechnol.* 14, 1685–1691.
- Wawer, C., Ruggeberg, H., Meyer, G., Muyzer, G., 1995. A simple and rapid electrophoresis method to detect sequence variation in PCR-amplified DNA fragments. *Nucleic Acid Res.* 23 (23), 4928–4929.

TOXICOLOGY

**An international journal concerned with
the effects of chemicals on living systems
and immunotoxicology**



**Univ. of Minn.
Bio-Medical
Library**

05 05 00

ELSEVIER

Special Issue

Festschrift dedicated to Professor Dr. K.J. Netter

Yeast microarrays for genome wide parallel genetic and gene expression analysis

DEVAL A. LASHKARI*†, JOSEPH L. DERISI‡, JOHN H. MCCUSKER§, ALLEN F. NAMATH‡, CRISTL GENTILE§, SEUNG Y. HWANG‡, PATRICK O. BROWN‡, AND RONALD W. DAVIS*‡

Departments of *Genetics and ‡Biochemistry, Stanford University, Stanford, CA 94305; and §Department of Microbiology, Duke University, Durham, NC 27710

Contributed by Ronald W. Davis, September 2, 1997

ABSTRACT We have developed high-density DNA microarrays of yeast ORFs. These microarrays can monitor hybridization to ORFs for applications such as quantitative differential gene expression analysis and screening for sequence polymorphisms. Automated scripts retrieved sequence information from public databases to locate predicted ORFs and select appropriate primers for amplification. The primers were used to amplify yeast ORFs in 96-well plates, and the resulting products were arrayed using an automated microarraying device. Arrays containing up to 2,479 yeast ORFs were printed on a single slide. The hybridization of fluorescently labeled samples to the array were detected and quantitated with a laser confocal scanning microscope. Applications of the microarrays are shown for genetic and gene expression analysis at the whole genome level.

The genome sequencing projects have generated and will continue to generate enormous amounts of sequence data. The genomes of *Saccharomyces cerevisiae*, *Haemophilus influenzae* (1), *Mycoplasma genitalium* (2), and *Methanococcus jannaschii* (3) have been completely sequenced. Other model organisms have had substantial portions of their genomes sequenced as well including the nematode *Caenorhabditis elegans* (4) and the small flowering plant *Arabidopsis thaliana* (5). Given this ever-increasing amount of sequence information, new strategies are necessary to efficiently pursue the next phase of the genome projects—the elucidation of gene expression patterns and gene product function on a whole genome scale.

One important use of genome sequence data is to attempt to identify the functions of predicted ORFs within the genome. Many of the ORFs identified in the yeast genome sequence were not identified in decades of genetic studies and have no significant homology to previously identified sequences in the database. In addition, even in cases where ORFs have significant homology to sequences in the database, or have known sequence motifs (e.g., protein kinase), this is not sufficient to determine the actual biological role of the gene product. Experimental analysis must be performed to thoroughly understand the biological function of a given ORF's product. Model organisms, such as *S. cerevisiae*, will be extremely important in improving our understanding of other more complex and less manipulable organisms.

To examine in detail the functional role of individual ORFs and relationships between genes at the expression level, this work describes the use of genome sequence information to study large numbers of genes efficiently and systematically. The procedure was as follows. (i) Software scripts scanned annotated sequence information from public databases for predicted ORFs. (ii) The start and stop position of each identified ORF was extracted automatically, along with the sequence data of the ORF and 200

bases flanking either side. (iii) These data were used to automatically select PCR primers that would amplify the ORF. (iv) The primer sequences were automatically input into the automated multiplex oligonucleotide synthesizer (6). (v) The oligonucleotides were synthesized in 96-well format, and (vi) used in 96-well format to amplify the desired ORFs from a genomic DNA template. (vii) The products were arrayed using a high-density DNA arrayer (7–10). The gene arrays can be used for hybridization with a variety of labeled products such as cDNA for gene expression analysis or genomic DNA for strain comparisons, and genomic mismatch scanning purified DNA for genotyping (11).

METHODS

Script Design. All scripts were written in UNIX Tool Command Language. Annotated sequence information from GenBank was extracted into one file containing the complete nucleotide sequence of a single chromosome. A second file contained the assigned ORF name followed by the start and stop positions of that ORF. The actual sequence contained within the specified range, along with 200 bases of sequence flanking both sides, was extracted and input into the primer selection program PRIMER 0.5 (Whitehead Institute, Boston). Primers were designed so as to allow amplification of entire ORFs. The selected primer sequences were read by the 96-well automated multiplex oligonucleotide synthesizer instrument for primer synthesis. The forward and reverse primers were synthesized in two separate 96-well plates in corresponding wells. All primers were synthesized on a 20-nmol scale.

ORF Amplification and Purification. Genomic DNA was isolated as described (12) and used as template for the amplification reactions. Each PCR was done in a total volume of 100 μ l. A total of 0.2 μ M each of forward and reverse primers were aliquoted into a 96-well PCR plate (Robbins Scientific, Sunnyvale, CA); a master mix containing 0.24 mM each dNTP, 10 mM Tris (pH 8.5), 50 mM MgCl₂, 2.5 units Taq polymerase, and 10 ng of template was added to the primers, and the entire mix was thermal cycled for 30 cycles as follows: 15 min at 94°C, 15 min at 54°C, and 30 min at 72°C. Products were ethanol precipitated in polystyrene v-bottom 96-well plates (Costar). All samples were dried and stored at –20°C.

Arraying Procedure and Processing. Microarrays were made as described (8).

A custom built arraying robot was used to print batches of 48 slides. The robot utilizes four printing tips which simultaneously pick up \sim 1 μ l of solution from 96-well microtiter plates. After printing, the microarrays were rehydrated for 30 sec in a humid chamber and then snap dried for 2 sec on a hot plate (100°C). The DNA was then UV crosslinked to the surface by subjecting the slides to 60 millijoules of energy. The rest of the poly-L-lysine surface was blocked by a 15-min incubation in a solution of 70 mM succinic anhydride dissolved in a solution consisting of 315 ml of 1-methyl-2-pyrrolidinone (Aldrich) and 35 ml of 1 M boric acid (pH 8.0). Directly after the blocking reaction, the bound DNA was denatured by a 2-min incubation in distilled water at \sim 95°C.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

© 1997 by The National Academy of Sciences 0027-8424/97/9413057-6\$2.00/0
PNAS is available online at <http://www.pnas.org>.

Abbreviation: YEP, yeast extract/peptone.

†To whom reprint requests should be sent at the present address: Synteni, Inc., 6519 Dumbarton Circle, Fremont, CA 94555.

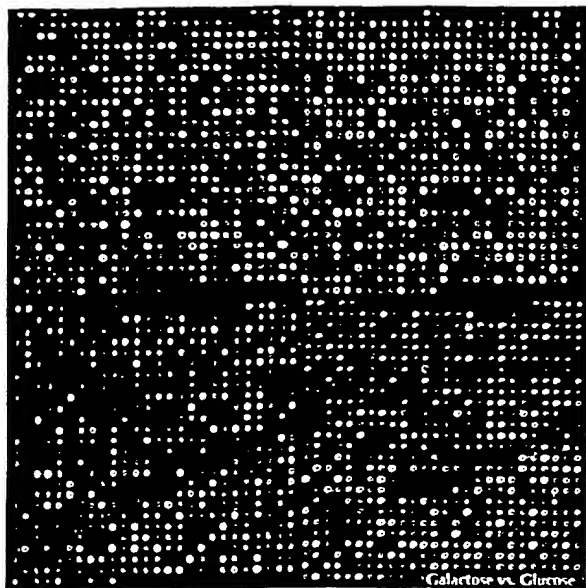


FIG. 1. Two-color fluorescent scan of a yeast microarray containing 2,479 elements (ORFs). The center-to-center distance between elements is 345 μm . A probe mixture consisting of cDNA from yeast extract/peptone (YEP) galactose (green pseudocolor) and YEP glucose (red pseudocolor) grown yeast cultures was hybridized to the array. Intensity per element corresponds to ORF expression, and pseudocolor per element corresponds to relative ORF expression between the two cultures.

The slides were then transferred into a bath of 100% ethanol at room temperature.

Probe Preparation: cDNA. Yeast cultures (100 ml) were grown to ~ 1 OD₆₀₀ and total RNA was isolated as described (13). Up to 500 μg total RNA was used to isolate mRNA (Qiagen, Chatsworth, CA). Oligo(dT)20 (5 μg) was added and annealed to 2 μg of mRNA by heating the reaction to 70°C for 10 min and quick chilling on ice, plus 2 μl SuperScript II (200 units/ μl) (Life Technologies, Gaithersburg, MD), 0.6 μl 50 \times dNTP mix (final concentrations were 500 μM dATP, dCTP, dGTP, and 200 μM dTTP), 6 μl 5 \times reaction buffer, and 60 μM Cy3-dUTP or Cy5-dUTP (Amersham). Reactions were carried out at 42°C for 2 h, after which the mRNA was degraded by the addition of 0.3 μl 5 M NaOH and 0.3 μl 100 mM EDTA and heating to 65°C for 10 min. The sample was then diluted to 500 μl with TE and concentrated using a Microcon-30 (Amicon) to 10 μl .

Probe Preparation: Genomic DNA. Fluorescent DNA was prepared from total genomic DNA as follows: 1 μg of random nonamer oligonucleotides was added to 2.5 μg of genomic DNA. This mixture was boiled for 2 min and then chilled on ice. A reaction mixture containing dNTPs (25 μM dATP, dCTP, dGTP, 10 μM dTTP, and 40 μM Cy3-dUTP or Cy5-dUTP) reaction buffer (New England Biolabs), and 20 units exonuclease free Klenow enzyme (United States Biochemical) was added, and the reaction was incubated at 37°C for 2 h. The sample was then diluted to 500 μl with TE and concentrated using a Microcon-30 (Amicon) to 10 μl .

Hybridization. Purified, labeled probe was resuspended in 11 μl of 3.5 \times SSC containing 10 μg *Escherichia coli* tRNA, and 0.3% SDS. The sample was then heated for 2 min in boiling water, cooled rapidly to room temperature, and applied to the array. The array was placed in a sealed, humidified, hybridization chamber. Hybridization was carried out for 10 h in a 62°C water bath, after which the arrays were washed immediately in 2 \times SSC/0.2% SDS. A second wash was performed in 0.1 \times SSC.

Analysis and Quantitation. Arrays were scanned on a scanning laser fluorescence microscope developed by Steve Smith with software written by Noam Ziv (Stanford Univer-

sity). A separate scan was done for each of the two fluorophores used. The images were then combined for analysis. A bounding box, fitted to the size of the DNA spots, was placed over each array element. The average fluorescent intensity was calculated by summing the intensities of each pixel present in a bounding box and then dividing by the total number of pixels. Local area background was calculated for each array element by determining the average fluorescent intensity at the edge of the bounding box. To normalize for fluorophore-specific variation, control spots containing yeast genomic DNA were applied to each quadrant during the arraying process. These elements were quantitated and the ratios of the signals were determined. These ratios were then used to normalize the photomultiplier sensitivity settings such that the ratios of the fluorescence of the genomic DNA spots were close to a value of 1.0. The average signal intensity at any given spot was regarded as significant if it was at least two standard deviations above background. Each experiment was conducted in duplicate, with the fluorophores representing each channel reversed. The ratios presented here are the average of the two experiments, except in the case in which the signal for the element in question was below the reliability threshold. The reliability threshold also determined the dynamic range of the experiment. For all of the experiments presented, the average dynamic range was ~ 1 to 100. In the case where the fluorescence from a very bright spot saturates the detector, differential ratios will, in general, be underestimated. This can be compensated for by scanning at a lower overall sensitivity.

RESULTS

The accumulation of sequence information from model organisms presents an enormous opportunity and challenge to understand the biological function of many previously uncharacterized genes. To do this accurately and efficiently, a directed strategy was developed that enables the monitoring of multiple genes simultaneously. Microarraying technology provides a method by which DNA can be attached to a glass surface in a high-density format (8). In practice, it is possible to array over 6,000 elements in an area less than 1.8 cm^2 . Given that the yeast genome consists of $\sim 6,100$ ORFs, the entire set of yeast genes can be spotted onto a single glass slide.

With this capability and the availability of the entire sequence of the yeast genome, our strategy was to use a directed approach for generating the complete genome array. This procedure involved synthesizing a pair of oligonucleotide primers to amplify each ORF. The PCR product containing each gene of interest was arrayed onto glass and used, for example, as probe for monitoring gene expression levels by hybridizing to the array labeled cDNA generated from isolated mRNA of a culture grown under any experimental condition.

Primer Selection and Synthesis. The primer selection was fully automated using Tool Command Language scripts and PRIMER 0.5. (Whitehead). Primer pairs were automatically selected successfully for >99% of the ORFs tested. Primer sequences can thus be selected rapidly with minimal manual processing. A complete set of forward and reverse primers were selected initially for each ORF on chromosomes I, II, III, V, VI, VIII, IX, X, and XI. Primers for a representative set of ORFs (15% coverage) were chosen for the remaining chromosomes. With the release of the entire yeast genome sequence, the complete set of primers has now been selected.

Because each ORF requires a unique pair of synthetic primers, a total of approximately 12,200 oligonucleotides will be required to individually amplify each target. This costly component was addressed with the automated multiplex oligonucleotide synthesizer (6) which efficiently synthesizes primers in a 96-well format. Each primer, synthesized on a 20-nmol scale, provides enough material for 100 amplification reactions, whereas a given PCR product provides enough material to generate an element on

Table 1. Heat shock vs. control expression data

Ratio of gene expression		ORF	Gene	Description
Control	Heat			
2.3	2.2	YLR142	PUT1	Proline oxidase
	2.0	YOL140	ARG8	Acetylornithine aminotransferase
		YGL148	ARO2	Chorismate synthase
	36.0	YFL014	HSP12	Heat shock protein
	27.4	YBR072	HSP26	Heat shock protein
	6.7	YBR054	YRO2	Similarity to HSP30 heat shock protein Yrolp
	3.4	YCR021	HSP30	Heat shock protein
	2.6	YER103	SSA4	Heat shock protein
	2.5	YLR259	HSP60	Mitochondrial heat shock protein HSP60
	2.1	YBR169	SSE2	Heat shock protein of the HSP70 family
	1.7	YBL075	SSA3	Cytoplasmic heat shock protein
	1.4	YPL240	HSP82	Heat shock protein
	1.4	YDR258	HSP78	Mitochondrial heat shock protein of clpb family of ATP-dependent proteases
	1.0	YNL007	SIS1	Heat shock protein
	1.1	YEL030		70-kDa heat shock protein
1.9		YHR064		Heat shock protein
2.6	1.3	YBL008	HIR1	Histone transcription regulator
		YBL002	HTB2	Histone H2B.2
	3.3	YBL003	HTA2	Histone H2A.2
	3.3	YBR010	HHT1	Histone H3
	3.9	YBR009	HHF1	Histone H4
	2.4	YDR343	HXT6	High-affinity hexose transporter
	2.1	YHR092	HXT4	Moderate- to low-affinity glucose transporter
	3.6	YAR071	PHO11	Secreted acid phosphatase, 56 kDa isozyme
	2.3	YLR096	KIN2	Ser/Thr protein kinase
	2.5	YER102	RPS8B	Ribosomal protein S8.e
	2.6	YBR181	RPS101	Ribosomal protein S6.e
	2.6	YCR031	CRY1	40S ribosomal protein S14.e
	2.7	YLR441	RP10A	Ribosomal protein S3.a.e
	2.8	YHR141	RPL41B	Ribosomal protein L36.a.e
	2.8	YBL072	RPS8A	Ribosomal protein S8.e
2.8		YHL015	URP2	Ribosomal protein
		YBR191	URP1	Ribosomal protein L21.e
	3.1	YLR340	RPLA0	Acidic Ribosomal protein L10.e
	3.3	YGL123	SUP44	Ribosomal protein
	5.8	YLR194		Hypothetical protein

500–1,000 arrays. Thus, a single primer pair provides enough starting material for up to ~50,000 arrays.

Primers were synthesized to amplify yeast ORFs. Primer synthesis had a failure rate of <1% in over 18 plates of synthesis as determined by standard trityl analysis (6). The success rate of the PCR amplifications using the primer pairs was 94% based on agarose gel analysis of each PCR. The purified PCR products were used to generate arrays. Two versions of the arrays were created for the experimental results presented here. The first array contained 2,287 elements and the second array batch contained 2,479 elements.

Genome Arrays. The amplified ORFs were arrayed onto glass at a spacing of 345 microns (Fig. 1). The high-density spacing of DNA samples allows the hybridization volumes to be minimized—volumes are a maximum of 10 μ l. The labeled probe can thus be maintained at relatively high concentrations, making 1–2 μ g of mRNA sufficient for analysis. This also obviates the need for a subsequent amplification step and thus avoids the risk of altering the relative ratios of different cDNA species in the sample.

Genetic Analysis: Genomic Comparison of Unrelated Strains. Microarrays allow efficient comparison of the genomes of different strains. Genomic DNA from Y55, an *S. cerevisiae* strain divergent from the reference strain S288c, was randomly labeled with Cy3-dUTP and hybridized simultaneously with the S288c DNA labeled with Cy5-dUTP. When a comparison between the hybridization of the DNA from the two strains was done, several

elements gave relatively little or no signal above background from the Cy3 channel (data not shown). These include SGE1, ASP3A-D, YLR156, YLR159, YLR161, ENA2 (YDR039 is ENA2), and YCR105. These results imply that the regions containing these genes are extremely divergent, or all together deleted from the strain. Subsequent attempts to generate PCR products from SGE1, ENA2, and ASP3A using Y55 DNA failed. This result supports the conclusion that these genes are likely to be missing from the Y55 genome. It is interesting to note that at least two of the regions absent in the Y55 genome have been previously shown or suggested to be deleted in mutant laboratory strains (14–16). In particular, the Asp-3 region appears to be highly prone to being deleted (15, 16).

These results indicate that gene arrays can be used to efficiently screen different strains of an organism for large deletion polymorphisms. A single hybridization and scan will reveal differences based on differential hybridization to particular elements. It is reasonable to suppose that an equivalent number of genes are present in the Y55 genome and absent in the S288c genome. This result should be viewed as a minimum estimate of the deletion polymorphisms that exist between these two unrelated strains as intergenic deletions or small intragenic deletions would not be detected because considerable hybridizing material would be remain. Sequence polymorphisms, such as deletions, are present in populations of every species and must at some level affect phenotype. One of the challenges of the genome era will be to critically examine sequence polymorphisms that exist in the natural gene pool relative to the reference genome sequence.

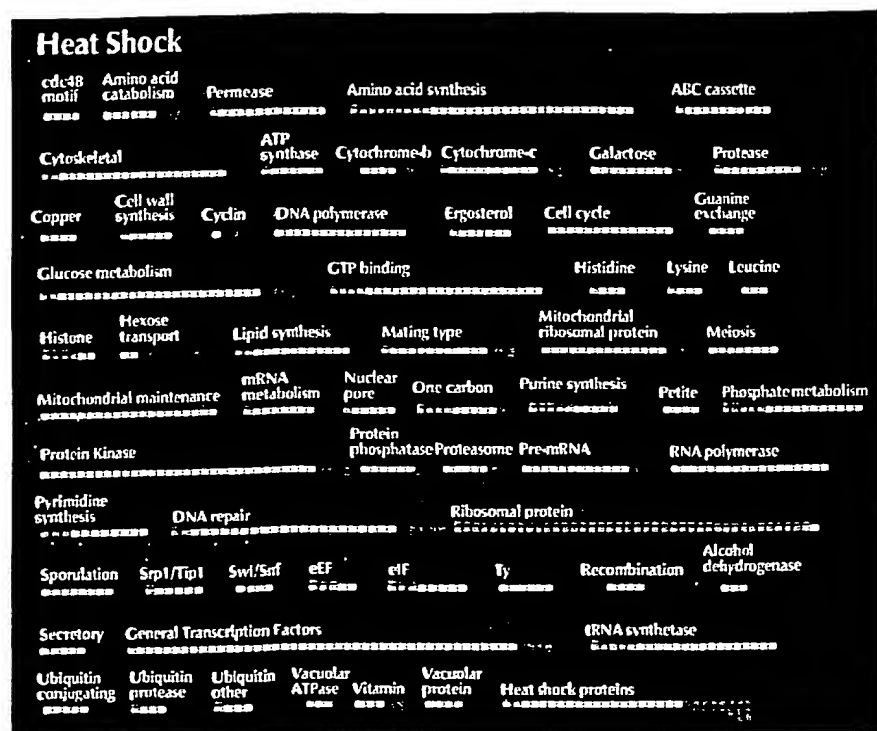


FIG. 2. ORF categories displaying differential expression between heat shocked and untreated cultures. Bars within categories correspond to individual ORFs. Green shaded bars correspond to relative increases in ORF expression under 25°C growth conditions. Red shaded bars correspond to relative increases in ORF expression under 39°C growth conditions.

Gene Expression Analysis. The arrays were used to examine gene expression in yeast grown under a variety of different conditions. Expression analysis is an ideal application of these arrays because a single hybridization provides quantitative expres-

sion data for thousands of genes. To better understand results for genes of known function, ORFs were placed in biologically relevant categories on the basis of function (e.g., amino acid catabolic genes) and/or pathways (e.g., the histidine biosynthesis pathway).

Table 2. Cold shock vs. control expression data

Ratio of gene expression		ORF	Gene	Description
Control	Cold			
	3.3	YOR153	PDR5	Pleiotropic drug resistance protein
2.4		YCR012	PGK1	Phosphoglycerate kinase
2.9		YCL040	GLK1	Aldohexose specific glucokinase
	1.4	YHR064		Heat shock protein
2.0		YJL034	KAR2	Nuclear fusion protein
2.1		YDR258	HSP78	Mitochondrial heat shock protein of clpb family of ATP-dependent proteases
2.2		YLL039	UBI4	Ubiquitin precursor
2.7		YLL026	HSP104	Heat shock protein
3.1		YER103	SSA4	Heat shock protein
3.3		YBR126	TPS1	α , α -Trehalose-phosphate synthase (UDP-forming)
3.8		YPL240	HSP82	Heat shock protein
7.9		YBR054	YRO2	Similarity to HSP30 heat shock protein Yro1p
7.9		YBR072	HSP26	Heat shock protein
16.5		YCR021	HSP30	Heat shock protein
1.8		YDR343	HXT6	High-affinity hexose transporter
2.1		YHR096	HXT5	Putative hexose transporter
2.4		YFR053	HXK1	Hexokinase I
2.8		YHR092	HXT4	Moderate- to low-affinity glucose transporter
3.4		YHR094	HXT1	Low-affinity hexose (glucose) transporter
	2.3	YHR089	GAR1	Nucleolar rRNA processing protein
	1.7	YLR048	NAB1B	40S ribosomal protein p40 homolog b
	1.7	YLR441	RP10A	Ribosomal protein S3a.e
	1.7	YLL045	RPL4B	Ribosomal protein L7a.e.B
	1.6	YLR029	RPL13A	Ribosomal protein L15.e
	1.6	YGL123	SUP44	Ribosomal protein
	3.1	YBR067	TIP1	Cold- and heat-shock-induced protein of the Srp1/Tip1p family
	2.2	YER011	TIR1	Cold-shock-induced protein of the Tir1p, Tip1p family
	2.0	YCR058		Hypothetical protein
	4.2	YKL102		Hypothetical protein

Table 3. Glucose vs. galactose expression data

Ratio of gene expression		ORF	Gene	Description
Glucose	Galactose			
2.1		YHR018	ARG4	Arginosuccinate lyase
3.5		YPR035	GLN1	Glutamate-ammonia ligase
2.8		YML116	ATR1	Aminotriazole and 4-nitroquinoline resistance protein
2.0		YMR303	ADH2	Alcohol dehydrogenase II
3.7		YBR145	ADH5	Alcohol dehydrogenase V
	3.2	YBL030	AAC2	ADP, ATP carrier protein 2
	2.9	YBR085	AAC3	ADP, ATP carrier protein
	2.7	YDR298	ATP5	H ⁺ -transporting ATP synthase δ chain precursor
	2.5	YBR039	ATP3	H ⁺ -transporting ATP synthase γ chain precursor
	5.5	YML054	CYB2	Lactate dehydrogenase cytochrome <i>b2</i>
	3.4	YML054	CYB2	Lactate dehydrogenase cytochrome <i>b2</i>
	2.3	YKL150	MCR1	Cytochrome- <i>b5</i> reductase
	4.2	YBL045	COR1	Ubiquinol-cytochrome <i>c</i> reductase 44K core protein
	3.5	YDL067	COX9	Cytochrome <i>c</i> oxidase chain VIIA
	2.7	YLR038	COX12	Cytochrome <i>c</i> oxidase, subunit VIB
	2.6	YHR051	COX6	Cytochrome <i>c</i> oxidase subunit VI
	2.4	YLR395	COX8	Cytochrome <i>c</i> oxidase chain VIII
	2.3	YFR033	QCR6	Ubiquinol-cytochrome <i>c</i> reductase 17K protein
	23.7	YLR081	GAL2	Galactose (and glucose) permease
	21.9	YBR018	GAL7	UDP-glucose-hexose-1-phosphate uridylyltransferase
	21.8	YBR020	GAL1	Galactokinase
	19.5	YBR019	GAL10	UDP-glucose 4-epimerase
	14.7	YLR081	GAL2	Galactose (and glucose) permease
	8.6	YDR009	GAL3	Galactokinase
	3.0	YML051	GAL80(1)	Negative regulator for expression of galactose-induced genes
	2.8	YML051	GAL80(2)	Negative regulator for expression of galactose-induced genes
2.7		YER055	HIS1	ATP phosphoribosyltransferase
3.4		YBR248	HIS7	Glutamine amidotransferase/cyclase
				Phosphoribosyl-AMP cyclohydrolase/phosphoribosyl-ATP pyrophosphatase/histidinol dehydrogenase
7.4		YCL030	HIS4	
5.8		YKR080	MTD1	Methylenetetrahydrofolate dehydrogenase (NAD ⁺)
6.0		YDR019	GCV1	Glycine decarboxylase T subunit
6.1		YLR058	SHM2	Serine hydroxymethyltransferase
	8.1	YML123	PHO84	High-affinity inorganic phosphate/H ⁺ symporter
3.5		YDR408	ADE8	Phosphoribosylglycinamide formyltransferase (GART)
3.6		YDR408	ADE8	Phosphoribosylglycinamide formyltransferase (GART)
4.4		YAR015	ADE1	Phosphoribosylamidoimidazole-succinocarboxamide synthase
5.6		YMR300	ADE4	Amidophosphoribosyltransferase
5.6		YOR128	ADE2	Phosphoribosylaminoimidazole carboxylase
6.0		YGL234	ADE5,7	Phosphoribosylamine-glycine ligase and phosphoribosylformylglycinamide cyclo-ligase
	6.3	YBL015	ACH1	Acetyl-CoA hydrolase

Heat Shock Results. A log phase culture growing in YEP/dextrose medium at 25°C was split in half. One half of the culture remained at 25°C whereas the other half of the culture was shifted to 39°C. mRNA was isolated from both cultures 1 h after heat shock for comparison on microarrays and, although this time point is not optimal for measuring induction of heat shock mRNAs (17), many known heat shock genes exhibited considerable induction at this time point (Table 1; Fig. 2). Down-regulation of genes in the ribosomal protein and histone gene categories was also observed. Differential expression between the heat-shocked culture and the control was also observed for many other genes. Genes in many categories, such as amino acid catabolism and amino acid synthesis, exhibited a mixed response with some genes showing little or no differential expression and other genes showing a significant increase or decrease in gene expression in response to heat shock (Table 1; Fig. 2).

Cold Shock Results. A log phase culture growing in YEP/dextrose medium at 37°C was split in half. One half of the culture remained at 37°C while the other half of the culture was shifted to 18°C. mRNA was isolated from both cultures 1 h after cold shock for comparison on microarrays. As expected,

two known cold shock genes (TIP1, TIR1) were expressed at a significantly higher level in the cold-shocked culture. Genes in other functional categories, such as glucose metabolism and heat shock displayed a mixed response with expression of some genes being unaffected and other genes exhibiting significant up- or down-regulation in response to cold shock (Table 2).

Steady-State Galactose vs. Glucose Results. mRNA was isolated from steady-state log phase YEP galactose and YEP glucose grown cultures for comparison on the microarrays. As expected, the GAL genes were expressed at a much higher level in the galactose culture. Many genes were differentially expressed in these cultures that were not *a priori* expected to exhibit differential expression. For example, some genes in the amino acid catabolic category were up-regulated in the galactose culture whereas genes in the one-carbon metabolism and purine categories were largely or entirely down-regulated in the galactose culture (Table 3). Genes in other categories, such as amino acid synthesis, abc transporter, cytochrome *c*, and cytochrome *b*, exhibited mixed responses; some genes in a category showed little or no obvious differential expression whereas other genes in the same category showed significant differential expression in the galactose and glucose cultures.

DISCUSSION

The results of these experiments show that many genes are differentially expressed under the three environmental conditions described here. The expected and predicted changes in gene expression, such as HSP12 in the heat-shocked culture, TIP1 in the cold-shocked culture, and GAL2 in the steady-state galactose culture, were observed in every case. However, in addition to the expected changes in gene expression, significant differential expression was also observed for many other genes that would not, *a priori*, be expected to be differentially expressed. For example, expression of PHO11 decreased and expression of YLR194, KIN2, and HXT6 increased in the heat shocked culture. Expression of MST1 and APE3 decreased and expression of PDR5 and GAR1 increased in the cold-shocked culture. In addition, ADE4 and SER2 were expressed at reduced levels whereas PHO84 and ACH1 were expressed at higher levels in cells grown in galactose compared with cells grown in glucose. Differential expression of these and many other genes was specific to one of these three environmental conditions.

Many other genes were found to be differentially expressed under more than one condition. When differentially expressed genes in cold- and heat-shocked cultures were compared, 30 genes were found in common. Of these 30 genes, 28 showed inverse expression (i.e., increased expression under one condition and decreased expression under the other condition). Two genes, YCR058 and YKL102, showed elevated expression in response to both cold and heat shock. Fifteen genes were found to be differentially expressed in both the heat-shocked and steady-state galactose cultures: 9 genes showed increased expression and 5 showed decreased expression under both conditions. Twenty genes were differentially expressed in both the cold-shocked and steady-state galactose cultures: 8 genes showed decreased expression and 5 genes showed increased expression under both conditions. Six genes showed increased expression in the galactose culture and decreased expression in the cold shocked culture. One gene (ODP1) showed increased expression in both the cold-shocked and steady-state galactose cultures.

Gene expression is affected in a global fashion when environmental conditions are changed and both expected and unexpected genes are affected. There is also overlap in the genes that are differentially expressed under quite different environmental conditions. These results can be rationalized by considering the high degree of cross-pathway regulation in yeast. For example, there is evidence for cross-pathway regulation between (i) carbon and nitrogen metabolism (18), (ii) phosphate and sulfate metabolism (19), and (iii) purine, phosphate, and amino acid metabolism (20–24). There are also examples of the interaction of general and specific transcription factors (25, 26). Finally, within the broad class of amino acid biosynthetic genes, there is evidence for amino acid specific regulation of some genes, regulation via general control for other genes, and regulation via both specific and general control for other genes (22, 27–30).

Cross-pathway regulation arises from the complex structure of promoters. Virtually all promoters contain sites for multiple transcription factors and, therefore, virtually all genes are subject to combinatorial regulation. For example, the HIS4 promoter contains binding sites for GCN4 (the general amino acid control transcription factor), PHO2/BAS2 (a transcriptional regulator of phosphatase and purine biosynthetic genes), and BAS1 (a transcriptional regulator of purine biosynthetic genes) (31). It is likely that the complex effects on gene expression described in this work are a direct consequence of the combinatorial regulation of gene expression.

These findings illustrate the power of the highly parallel whole genome approach when examining gene expression. The global effects of environmental change on gene expression can now be directly visualized. It is clear that determining the mechanism(s) and the functional role of the dramatic global effects on gene

expression in different environments will be a significant challenge. The era of whole genome analysis will, ultimately, allow researchers to switch from the very focused single gene/promoter view of gene expression and instead view the cell more as a large complex network of gene regulatory pathways.

With the entire sequence of this model organism known, new approaches have been developed that allow for genome wide analyses (32, 33) of gene function. The genome microarrays represent a novel tool for genetic and expression analysis of the yeast genome. This pilot study uses arrays containing >35% of the yeast ORFs and it is clear that the entire set of ORFs from the yeast genome can be arrayed using the directed primer based strategy detailed here. Recent advances in arraying technology will allow all 6,100 ORFs to be arrayed in an area of less than 1.8 cm². Furthermore, as the technology improves, detection limits will allow less than 500 ng of starting mRNA material to be used for making probe.

The genome arrays provide for a robust, fully automated approach toward examining genome structure and gene function. They allow for comparisons between different genomes as well as a detailed study of gene expression at the global level. This research will help to elucidate relationships between genes and allow the researcher to understand gene function by understanding expression patterns across the yeast genome.

Support was provided by National Institutes of Health Grant P01HG00205.

1. Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., *et al.* (1995) *Science* 269, 496–512.
2. Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., *et al.* (1995) *Science* 270, 397–403.
3. Bult, C. J., White, O., Olsen, G. J., Zhou, L., Fleischmann, R. D., *et al.* (1996) *Science* 273, 1058–1073.
4. Sulston, J., Du, Z., Thomas, K., Wilson, R., Hillier, L., *et al.* (1992) *Nature (London)* 356, 37.
5. Newman, T., de Bruijn, F. J., Green, P., Keegstra, K., Kende, H., *et al.* (1994) *Plant Physiol.* 106, 1241–1255.
6. Lashkari, D. A., Hunnicke-Smith, S. P., Norgren, R. M., Davis, R. W., & Brennan, T. (1995) *Proc. Natl. Acad. Sci. USA* 92, 7912–7915.
7. Schena, M., Shalon, D., Davis, R. W., & Brown, P. O. (1995) *Science* 270, 467–470.
8. Shalon, D., Smith, S., & Brown, P. O. (1996) *Genome Res.* 6, 639–645.
9. Heller, R. A., Schena, M., Chai, A., Shalon, D., Bedilion, T., Gilmore, J., Woolley, D. E., & Davis, R. W. (1997) *Proc. Natl. Acad. Sci. USA* 94, 2150–2155.
10. Derisi, J., Penland, L., Brown, P. O., Bittner, M. L., Meltzer, P. S., Ray, M., Chen, Y., Su Ya, & Trent, J. M. (1996) *Nat. Genet.* 14, 457–460.
11. Nelson, S. F., McCusker, J. H., Sander, M. A., Kee, Y., Modrich, P., & Brown P. O. (1993) *Nat. Genet.* 4, 11–18.
12. Hoffman, C. S., & Winston, F. (1989) *Gene* 84, 473–479.
13. Schmitt, M., Brown, T., & Trumpower, B. (1990) *Nucleic Acids Res.* 18, 3091.
14. Ehrenhofer-Murray, A. E., Wurgler, F. E., & Sengstag, C. (1994) *Mol. Gen. Genet.* 244, 287–294.
15. Kim, K.-W., Kamerud, J. Q., Livingston, D. M., & Roon, R. J. (1988) *J. Biol. Chem.* 263, 11948–11953.
16. Kim, K.-W., & Roon, R. J. (1984) *J. Bacteriol.* 157, 958–961.
17. Craig, E. A. (1992) in *The Molecular Biology of the Yeast Saccharomyces: Gene Expression*, eds. Jones, E. W., Pringle, J. R., & Broach, J. R. (Cold Spring Harbor Lab. Press, Plainview, NY), Vol. 2, pp. 501–537.
18. Dang, V. D., Bohn, C., Bolotin-Fukuhara, M., & Daignan-Fornier, B. (1996) *J. Bacteriol.* 178, 1842–1849.
19. O'Connell, K. F., & Baker, R. E. (1992) *Genetics* 132, 63–73.
20. Braus, G., Mosch, H. U., Vogel, K., Hinnen, A., & Hutter, R. (1989) *EMBO J.* 8, 939–945.
21. Mosch, H. U., Scheiber, B., Lahti, R., Mantsala, P., & Braus, G. H. (1991) *J. Biol. Chem.* 266, 20453–20456.
22. Mitchell, A. P., & Magasanik, B. (1984) *Mol. Cell. Biol.* 4, 2761–2773.
23. Daignan-Fornier, B., & Fink, G. R. (1992) *Proc. Natl. Acad. Sci. USA* 89, 6746–6750.
24. Tice-Baldwin, K., Fink, G. R., & Arndt, K. T. (1989) *Science* 246, 931–935.
25. Messenguy, F., & Dubois, E. (1993) *Mol. Cell. Biol.* 13, 2586–2592.
26. Devlin, C., Tice-Baldwin, K., Shore, D., & Arndt, K. T. (1991) *Mol. Cell. Biol.* 11, 3642–3651.
27. Magasanik, B. (1992) in *The Molecular and Cellular Biology of the Yeast Saccharomyces: Gene Expression*, eds. Jones, E. W., Pringle, J. R., & Broach, J. R. (Cold Spring Harbor Lab. Press, Plainview, NY), Vol. 2, pp. 283–317.
28. Hinnebusch, A. G. (1992) in *The Molecular and Cellular Biology of the Yeast Saccharomyces: Gene Expression*, eds. Jones, E. W., Pringle, J. R., & Broach, J. R. (Cold Spring Harbor Lab. Press, Plainview, NY), Vol. 2, pp. 319–414.
29. Brisco, P. R., & Kohlhaw, G. B. (1990) *J. Biol. Chem.* 265, 11667–11675.
30. O'Connell, K. F., Surdin-Kerjan, Y., & Baker, R. E. (1995) *Mol. Cell. Biol.* 15, 1879–1888.
31. Arndt, K. T., Styles, C., & Fink, G. R. (1987) *Science* 237, 874–880.
32. Smith, V., Chou, K. N., Lashkari, D., Botstein, D., & Brown, P. O. (1996) *Science* 274, 2069–2074.
33. Shoemaker, D. D., Lashkari, D. A., Morris, D., Mittman, M., & Davis, R. W. (1996) *Nat. Genet.* 14, 450–456.

- Fischer-Vize, *Science* 270, 1828 (1995).
35. T. C. James and S. C. Elgin, *Mol. Cell Biol.* 6, 3862 (1986); R. Paro and D. S. Hogness, *Proc. Natl. Acad. Sci. U.S.A.* 88, 263 (1991); B. Tschiersch *et al.*, *EMBO J.* 13, 3822 (1994); M. T. Madireddi *et al.*, *Cell* 87, 75 (1996); D. G. Stokes, K. D. Tartof, R. P. Perry, *Proc. Natl. Acad. Sci. U.S.A.* 93, 7137 (1996).
 36. P. M. Palosaari *et al.*, *J. Biol. Chem.* 266, 10750 (1991); A. Schmitz, K. H. Gantemann, J. Fiedler, E. Grund, R. Eichenlaub, *Appl. Environ. Microbiol.* 58, 4068 (1992); V. Sharma, K. Suvama, R. Megathanan, M. E. Hudspeth, *J. Bacteriol.* 174, 5057 (1992); M. Kanazawa *et al.*, *Enzyme Protein* 47, 9 (1993); Z. L. Boynton, G. N. Bennet, F. B. Rudolph, *J. Bacteriol.* 178, 3015 (1996).
 37. M. Ho *et al.*, *Cell* 77, 869 (1994).
 38. W. Hendriks *et al.*, *J. Cell Biochem.* 59, 418 (1995).
 39. We thank H. Skaletsky and F. Lewitter for help with

28 April 1997; accepted 9 September 1997

Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale

Joseph L. DeRisi, Vishwanath R. Iyer, Patrick O. Brown*

DNA microarrays containing virtually every gene of *Saccharomyces cerevisiae* were used to carry out a comprehensive investigation of the temporal program of gene expression accompanying the metabolic shift from fermentation to respiration. The expression profiles observed for genes with known metabolic functions pointed to features of the metabolic reprogramming that occur during the diauxic shift, and the expression patterns of many previously uncharacterized genes provided clues to their possible functions. The same DNA microarrays were also used to identify genes whose expression was affected by deletion of the transcriptional co-repressor *TUP1* or overexpression of the transcriptional activator *YAP1*. These results demonstrate the feasibility and utility of this approach to genomewide exploration of gene expression patterns.

using a simple robotic printing device (9). Cells from an exponentially growing culture of yeast were inoculated into fresh medium and grown at 30°C for 21 hours. After an initial 9 hours of growth, samples were harvested at seven successive 2-hour intervals, and mRNA was isolated (10). Fluorescently labeled cDNA was prepared by reverse transcription in the presence of Cy3(green)- or Cy5(red)-labeled deoxyuridine triphosphate (dUTP) (11) and then hybridized to the microarrays (12). To maximize the reliability with which changes in expression levels could be discerned, we labeled cDNA prepared from cells at each successive time point with Cy5, then mixed it with a Cy3-labeled "reference" cDNA sample prepared from cells harvested at the first interval after inoculation. In this experimental design, the relative fluorescence intensity measured for the Cy3 and Cy5 fluors at each array element provides a reliable measure of the relative abundance of the corresponding mRNA in the two cell populations (Fig. 1). Data from the series of seven samples (Fig. 2), consisting of more than 43,000 expression-ratio measurements, were organized into a database to facilitate efficient exploration and analysis of the results. This database is publicly available on the Internet (13).

During exponential growth in glucose-rich medium, the global pattern of gene expression was remarkably stable. Indeed, when gene expression patterns between the first two cell samples (harvested at a 2-hour interval) were compared, mRNA levels differed by a factor of 2 or more for only 19 genes (0.3%), and the largest of these differences was only 2.7-fold (14). However, as glucose was progressively depleted from the growth media during the course of the experiment, a marked change was seen in the global pattern of gene expression. mRNA levels for approximately 710 genes were induced by a factor of at least 2, and the mRNA levels for approximately 1030 genes declined by a factor of at least 2. Messenger RNA levels for 183 genes increased by a factor of at least 4, and mRNA levels for 203 genes diminished by a factor of at least 4. About half of these differentially expressed genes have no currently recognized function and are not yet named. Indeed, more than 400 of the differentially expressed genes have no apparent homology

The complete sequences of nearly a dozen microbial genomes are known, and in the next several years we expect to know the complete genome sequences of several metazoans, including the human genome. Defining the role of each gene in these genomes will be a formidable task, and understanding how the genome functions as a whole in the complex natural history of a living organism presents an even greater challenge.

Knowing when and where a gene is expressed often provides a strong clue as to its biological role. Conversely, the pattern of genes expressed in a cell can provide detailed information about its state. Although regulation of protein abundance in a cell is by no means accomplished solely by regulation of mRNA, virtually all differences in cell type or state are correlated with changes in the mRNA levels of many genes. This is fortuitous because the only specific reagent required to measure the abundance of the mRNA for a specific gene is a cDNA sequence. DNA microarrays, consisting of thousands of individual gene sequences printed in a high-density array on a glass microscope slide (1, 2), provide a practical and economical tool for studying gene expression on a very large scale (3–6).

Saccharomyces cerevisiae is an especially

favorable organism in which to conduct a systematic investigation of gene expression. The genes are easy to recognize in the genome sequence, cis regulatory elements are generally compact and close to the transcription units, much is already known about its genetic regulatory mechanisms, and a powerful set of tools is available for its analysis.

A recurring cycle in the natural history of yeast involves a shift from anaerobic (fermentation) to aerobic (respiration) metabolism. Inoculation of yeast into a medium rich in sugar is followed by rapid growth fueled by fermentation, with the production of ethanol. When the fermentable sugar is exhausted, the yeast cells turn to ethanol as a carbon source for aerobic growth. This switch from anaerobic growth to aerobic respiration upon depletion of glucose, referred to as the diauxic shift, is correlated with widespread changes in the expression of genes involved in fundamental cellular processes such as carbon metabolism, protein synthesis, and carbohydrate storage (7). We used DNA microarrays to characterize the changes in gene expression that take place during this process for nearly the entire genome, and to investigate the genetic circuitry that regulates and executes this program.

Yeast open reading frames (ORFs) were amplified by the polymerase chain reaction (PCR), with a commercially available set of primer pairs (8). DNA microarrays, containing approximately 6400 distinct DNA sequences, were printed onto glass slides by

Department of Biochemistry, Stanford University School of Medicine, Howard Hughes Medical Institute, Stanford, CA 94305–5428, USA.

*To whom correspondence should be addressed. E-mail: pbrown@crgm.stanford.edu

to any gene whose function is known (15). The responses of these previously uncharacterized genes to the diauxic shift therefore provides the first small clue to their possible roles.

The global view of changes in expression of genes with known functions provides a vivid picture of the way in which the cell adapts to a changing environment. Figure 3 shows a portion of the yeast metabolic pathways involved in carbon and energy metabolism. Mapping the changes we observed in the mRNAs encoding each enzyme onto this framework allowed us to infer the redirection in the flow of metabolites through this system. We observed large inductions of the genes coding for the enzymes aldehyde dehydrogenase (ALD2) and acetyl-coenzyme A (CoA) synthase (ACS1), which function together to convert the products of alcohol dehydrogenase into acetyl-CoA, which in turn is used to fuel the tricarboxylic acid (TCA) cycle and the glyoxylate cycle. The concomitant shutdown of transcription of the genes encoding pyruvate decarboxylase and induction of pyruvate carboxylase rechannels pyruvate away from acetaldehyde, and instead to oxalacetate, where it can serve to supply the TCA cycle and gluconeogenesis. Induction of the pivotal genes *PCK1*, encoding phosphoenolpyruvate carboxykinase, and *FBP1*, encoding fructose 1,6-bisphosphatase, switches the directions of two key irreversible steps in glycolysis, reversing the flow of metabolites along the reversible steps of the glycolytic pathway toward the essential biosynthetic precursor, glucose-6-phosphate. Induction of the genes coding for the trehalose synthase and glycogen synthase complexes promotes channeling of glucose-6-phosphate into these carbohydrate storage pathways.

Just as the changes in expression of genes encoding pivotal enzymes can provide insight into metabolic reprogramming, the behavior of large groups of functionally related genes can provide a broad view of the systematic way in which the yeast cell adapts to a changing environment (Fig. 4). Several classes of genes, such as cytochrome *c*-related genes and those involved in the TCA/glyoxylate cycle and carbohydrate storage, were coordinately induced by glucose exhaustion. In contrast, genes devoted to protein synthesis, including ribosomal proteins, tRNA synthetases, and translation, elongation, and initiation factors, exhibited a coordinated decrease in expression. More than 95% of ribosomal genes showed at least twofold decreases in expression during the diauxic shift (Fig. 4) (13). A noteworthy and illuminating exception was that the

genes encoding mitochondrial ribosomal genes were generally induced rather than repressed after glucose limitation, highlighting the requirement for mitochondrial biogenesis (13). As more is learned about the functions of every gene in the yeast genome, the ability to gain insight into a cell's response to a changing environment through its global gene expression patterns will become increasingly powerful.

Several distinct temporal patterns of expression could be recognized, and sets of genes could be grouped on the basis of the similarities in their expression patterns. The characterized members of each of these groups also shared important similarities in their functions. Moreover, in most cases, common regulatory mechanisms could be inferred for sets of genes with similar expression profiles. For example, seven genes showed a late induction profile, with mRNA levels increasing by more than ninefold at

the last timepoint but less than threefold at the preceding timepoint (Fig. 5B). All of these genes were known to be glucose-repressed, and five of the seven were previously noted to share a common upstream activating sequence (UAS), the carbon source response element (CSRE) (16–20). A search in the promoter regions of the remaining two genes, *ACR1* and *IDP2*, revealed that *ACR1*, a gene essential for ACS1 activity, also possessed a consensus CSRE motif, but interestingly, *IDP2* did not. A search of the entire yeast genome sequence for the consensus CSRE motif revealed only four additional candidate genes, none of which showed a similar induction.

Examples from additional groups of genes that shared expression profiles are illustrated in Fig. 5, C through F. The sequences upstream of the named genes in Fig. 5C all contain stress response elements (STRE), and with the exception

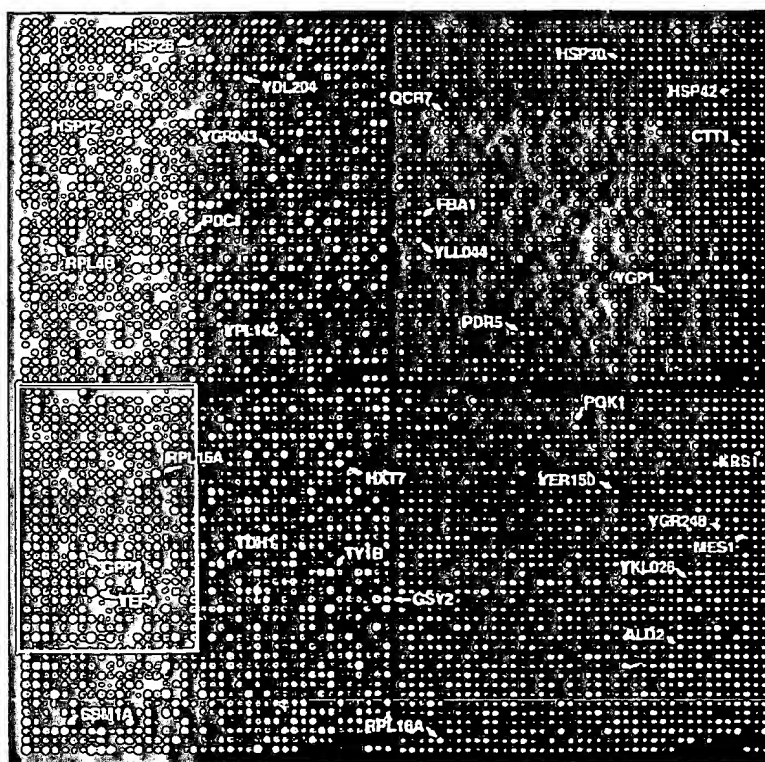


Fig. 1. Yeast genome microarray. The actual size of the microarray is 18 mm by 18 mm. The microarray was printed as described (9). This image was obtained with the same fluorescent scanning confocal microscope used to collect all the data we report (49). A fluorescently labeled cDNA probe was prepared from mRNA isolated from cells harvested shortly after inoculation (culture density of $<5 \times 10^6$ cells/ml and media glucose level of 19 g/liter) by reverse transcription in the presence of Cy3-dUTP. Similarly, a second probe was prepared from mRNA isolated from cells taken from the same culture 9.5 hours later (culture density of $\sim 2 \times 10^8$ cells/ml, with a glucose level of <0.2 g/liter) by reverse transcription in the presence of Cy5-dUTP. In this image, hybridization of the Cy3-dUTP-labeled cDNA (that is, mRNA expression at the initial timepoint) is represented as a green signal, and hybridization of Cy5-dUTP-labeled cDNA (that is, mRNA expression at 9.5 hours) is represented as a red signal. Thus, genes induced or repressed after the diauxic shift appear in this image as red and green spots, respectively. Genes expressed at roughly equal levels before and after the diauxic shift appear in this image as yellow spots.

of HSP42, have previously been shown to be controlled at least in part by these elements (21–24). Inspection of the sequences upstream of HSP42 and the two uncharacterized genes shown in Fig. 5C, YKL026c, a hypothetical protein with similarity to glutathione peroxidase, and YGR043c, a putative transaldolase, revealed that each of these genes also possess repeated upstream copies of the stress-responsive CCCCT motif. Of the 13 additional genes in the yeast genome that shared this expression profile [including HSP30, ALD2, OM45, and 10 uncharacterized ORFs (25)], nine contained one or more recognizable STRE sites in their upstream regions.

The heterotrimeric transcriptional activator complex HAP2,3,4 has been shown to be responsible for induction of several genes important for respiration (26–28). This complex binds a degenerate consensus sequence known as the CCAAT box (26). Computer analysis, using the consensus sequence TNRYTGGB (29), has suggested that a large number of genes involved in respiration may be specific targets of HAP2,3,4 (30). Indeed, a putative HAP2,3,4 binding site could be found in the sequences upstream of each of the seven cytochrome c-related genes that showed the greatest magnitude of induction (Fig. 5D). Of 12 additional cytochrome c-related genes that were induced, HAP2,3,4 binding sites were present in all but one. Significantly, we found that transcription of HAP4 itself was induced nearly ninefold concomitant with the diauxic shift.

Control of ribosomal protein biogenesis is mainly exerted at the transcriptional level, through the presence of a common upstream-activating element (UAS_{rp}) that is recognized by the Rap1 DNA-binding protein (31, 32). The expression profiles of seven ribosomal proteins are shown in Fig. 5F. A search of the sequences upstream of all seven genes revealed consensus Rap1-binding motifs (33). It has been suggested that declining Rap1 levels in the cell during starvation may be responsible for the decline in ribosomal protein gene expression (34). Indeed, we observed that the abundance of RAP1 mRNA diminished by 4.4-fold, at about the time of glucose exhaustion.

Of the 149 genes that encode known or putative transcription factors, only two, HAP4 and SIP4, were induced by a factor of more than threefold at the diauxic shift. SIP4 encodes a DNA-binding transcriptional activator that has been shown to interact with Snf1, the “master regulator” of glucose repression (35). The eightfold induction of SIP4 upon depletion of glucose strongly suggests a role in the induction of

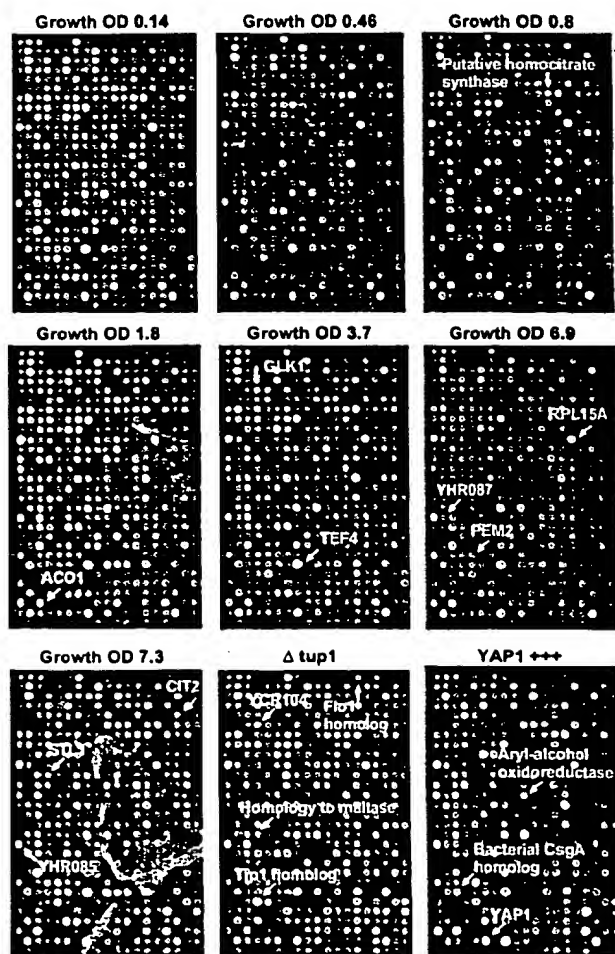
downstream genes at the diauxic shift.

Although most of the transcriptional responses that we observed were not previously known, the responses of many genes during the diauxic shift have been described. Comparison of the results we obtained by DNA microarray hybridization with previously reported results therefore provided a strong test of the sensitivity and accuracy of this approach. The expression patterns we observed for previously characterized genes showed almost perfect concordance with previously published results (36). Moreover, the differential expression measurements obtained by DNA microarray hybridization were reproducible in duplicate experiments. For example, the remarkable changes in gene expression between cells harvested immediately after inoculation and immediately after the diauxic shift (the first and sixth intervals in this time series) were measured in duplicate, independent DNA microarray hybridizations. The correlation coefficient for two complete sets of expression ratio measurements was 0.87, and for more than 95% of the genes, the expres-

sion ratios measured in these duplicate experiments differed by less than a factor of 2. However, in a few cases, there were discrepancies between our results and previous results, pointing to technical limitations that will need to be addressed as DNA microarray technology advances (37, 38). Despite the noted exceptions, the high concordance between the results we obtained in these experiments and those of previous studies provides confidence in the reliability and thoroughness of the survey.

The changes in gene expression during this diauxic shift are complex and involve integration of many kinds of information about the nutritional and metabolic state of the cell. The large number of genes whose expression is altered and the diversity of temporal expression profiles observed in this experiment highlight the challenge of understanding the underlying regulatory mechanisms. One approach to defining the contributions of individual regulatory genes to a complex program of this kind is to use DNA microarrays to identify genes whose expression is affected

Fig. 2. The section of the array indicated by the gray box in Fig. 1 is shown for each of the experiments described here. Representative genes are labeled. In each of the arrays used to analyze gene expression during the diauxic shift, red spots represent genes that were induced relative to the initial timepoint, and green spots represent genes that were repressed relative to the initial timepoint. In the arrays used to analyze the effects of the *tup1*Δ mutation and YAP1 overexpression, red spots represent genes whose expression was increased, and green spots represent genes whose expression was decreased by the genetic modification. Note that distinct sets of genes are induced and repressed in the different experiments. The complete images of each of these arrays can be viewed on the Internet (13). Cell density as measured by optical density (OD) at 600 nm was used to measure the growth of the culture.



by mutations in each putative regulatory gene. As a test of this strategy, we analyzed the genomewide changes in gene expression that result from deletion of the *TUP1* gene. Transcriptional repression of many genes by glucose requires the DNA-binding repressor

Mig1 and is mediated by recruiting the transcriptional co-repressors Tup1 and Cyc8/Ssn6 (39). Tup1 has also been implicated in repression of oxygen-regulated, mating-type-specific, and DNA-damage-inducible genes (40).

Wild-type yeast cells and cells bearing a deletion of the *TUP1* gene (*tup1Δ*) were grown in parallel cultures in rich medium containing glucose as the carbon source. Messenger RNA was isolated from exponentially growing cells from the two populations and used to prepare cDNA labeled with Cy3 (green) and Cy5 (red), respectively (11). The labeled probes were mixed and simultaneously hybridized to the microarray. Red spots on the microarray therefore represented genes whose transcription was induced in the *tup1Δ* strain, and thus presumably repressed by Tup1 (41). A representative section of the microarray (Fig. 2, bottom middle panel) illustrates that the genes whose expression was affected by the *tup1Δ* mutation, were, in general, distinct from those induced upon glucose exhaustion [complete images of all the arrays shown in Fig. 2 are available on the Internet (13)]. Nevertheless, 34 (10%) of the genes that were induced by a factor of at least 2 after the diauxic shift were similarly induced by deletion of *TUP1*, suggesting that these genes may be subject to *TUP1*-mediated repression by glucose. For example, *SUC2*, the gene encoding invertase, and all five hexose transporter genes that were induced during the course of the diauxic shift were similarly induced, in duplicate experiments, by the deletion of *TUP1*.

The set of genes affected by Tup1 in this experiment also included α -glucosidases, the mating-type-specific genes *MFA1* and *MFA2*, and the DNA damage-inducible *RNR2* and *RNR4*, as well as genes involved in flocculation and many genes of unknown function. The hybridization signal corresponding to expression of *TUP1* itself was also severely reduced because of the (incomplete) deletion of the transcription unit in the *tup1Δ* strain, providing a positive control in the experiment (42).

Many of the transcriptional targets of Tup1 fell into sets of genes with related biochemical functions. For instance, although only about 3% of all yeast genes appeared to be *TUP1*-repressed by a factor of more than 2 in duplicate experiments under these conditions, 6 of the 13 genes that have been implicated in flocculation (15) showed a reproducible increase in expression of at least twofold when *TUP1* was deleted. Another group of related genes that appeared to be subject to *TUP1* repression encodes the serine-rich cell wall mannoproteins, such as *Tip1* and *Tir1/Srp1* which are induced by cold shock and other stresses (43), and similar, serine-poor proteins, the seripauperins (44). Messenger RNA levels for 23 of the 26 genes in this group were reproducibly elevated by at least 2.5-fold in the *tup1Δ*

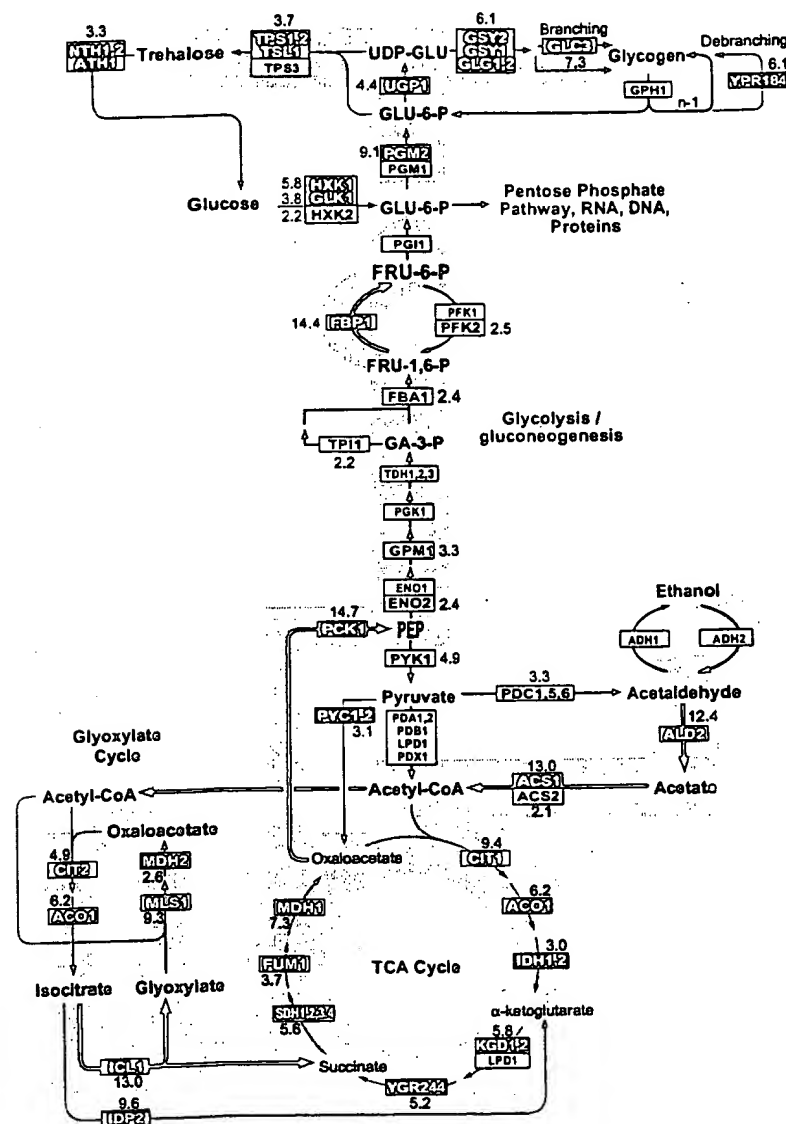


Fig. 3. Metabolic reprogramming inferred from global analysis of changes in gene expression. Only key metabolic intermediates are identified. The yeast genes encoding the enzymes that catalyze each step in this metabolic circuit are identified by name in the boxes. The genes encoding succinyl-CoA synthase and glycogen-debranching enzyme have not been explicitly identified, but the ORFs *YGR244* and *YPR184* show significant homology to known succinyl-CoA synthase and glycogen-debranching enzymes, respectively, and are therefore included in the corresponding steps in this figure. Red boxes with white lettering identify genes whose expression increases in the diauxic shift. Green boxes with dark green lettering identify genes whose expression diminishes in the diauxic shift. The magnitude of induction or repression is indicated for these genes. For multimeric enzyme complexes, such as succinate dehydrogenase, the indicated fold-induction represents an unweighted average of all the genes listed in the box. Black and white boxes indicate no significant differential expression (less than twofold). The direction of the arrows connecting reversible enzymatic steps indicate the direction of the flow of metabolic intermediates, inferred from the gene expression pattern, after the diauxic shift. Arrows representing steps catalyzed by genes whose expression was strongly induced are highlighted in red. The broad gray arrows represent major increases in the flow of metabolites after the diauxic shift, inferred from the indicated changes in gene expression.

strain, and 18 of these genes were induced by more than sevenfold when *TUP1* was deleted. In contrast, none of 83 genes that could be classified as putative regulators of the cell division cycle were induced more than twofold by deletion of *TUP1*. Thus, despite the diversity of the regulatory systems that employ Tup1, most of the genes that it regulates under these conditions fall into a limited number of distinct functional classes.

Because the microarray allows us to monitor expression of nearly every gene in yeast, we can, in principle, use this approach to identify all the transcriptional targets of a regulatory protein like Tup1. It is important to note, however, that in any single experiment of this kind we can only recognize those target genes that are normally repressed (or induced) under the conditions of the experiment. For instance, the experiment described here analyzed a MAT α strain in which *MFA1* and *MFA2*, the genes encoding the α -factor mating pheromone precursor, are normally repressed. In the isogenic *tup1 Δ* strain, these genes were inappropriately expressed, reflecting the role that Tup1 plays in their repression. Had we instead carried out this experiment with a MAT α strain (in which expression of *MFA1* and *MFA2* is not repressed), it would not have been possible to conclude anything regarding the role of Tup1 in the repression of these genes. Conversely, we cannot distinguish indirect effects of the chronic absence of Tup1 in the mutant strain from effects directly attributable to its participation in repressing the transcription of a gene.

Another simple route to modulating the activity of a regulatory factor is to overexpress the gene that encodes it. *YAP1* encodes a DNA-binding transcription factor belonging to the bZIP class of DNA-binding proteins. Overexpression of *YAP1* in yeast confers increased resistance to hydrogen peroxide, *o*-phenanthroline, heavy metals, and osmotic stress (45). We analyzed differential gene expression between a wild-type strain bearing a control plasmid and a strain with a plasmid expressing *YAP1* under the control of the strong *GAL1-10* promoter, both grown in galactose (that is, a condition that induces *YAP1* overexpression). Complementary DNA from the control and *YAP1* overexpressing strains, labeled with Cy3 and Cy5, respectively, was prepared from mRNA isolated from the two strains and hybridized to the microarray. Thus, red spots on the array represent genes that were induced in the strain overexpressing *YAP1*.

Of the 17 genes whose mRNA levels increased by more than threefold when

YAP1 was overexpressed in this way, five bear homology to aryl-alcohol oxidoreductases (Fig. 2 and Table 1). An additional four of the genes in this set also belong to the general class of dehydrogenases/oxidoreductases. Very little is known about the role of aryl-alcohol oxidoreductases in *S. cerevisiae*, but these enzymes have been isolated from ligninolytic fungi, in which they participate in coupled redox reactions, oxidizing aromatic, and aliphatic unsaturated alcohols to aldehydes with the production of hydrogen peroxide (46, 47). The fact that a remarkable fraction of the targets identified in this experiment belong to the same small, functional group of oxidoreductases suggests that these genes

might play an important protective role during oxidative stress. Transcription of a small number of genes was reduced in the strain overexpressing *Yap1*. Interestingly, many of these genes encode sugar permeases or enzymes involved in inositol metabolism.

We searched for *Yap1*-binding sites (TTACTAA or TGACTAA) in the sequences upstream of the target genes we identified (48). About two-thirds of the genes that were induced by more than threefold upon *Yap1* overexpression had one or more binding sites within 600 bases upstream of the start codon (Table 1), suggesting that they are directly regulated by *Yap1*. The absence of canonical *Yap1*-bind-

Fig. 4. Coordinated regulation of functionally related genes. The curves represent the average induction or repression ratios for all the genes in each indicated group. The total number of genes in each group was as follows: ribosomal proteins, 112; translation elongation and initiation factors, 25; tRNA synthetases (excluding mitochondrial synthetases), 17; glycogen and trehalose synthesis and degradation, 15; cytochrome c oxidase and reductase proteins, 19; and TCA- and glyoxylate-cycle enzymes, 24.

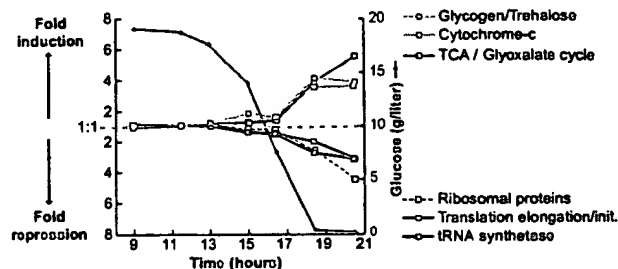


Table 1. Genes induced by *YAP1* overexpression. This list includes all the genes for which mRNA levels increased by more than twofold upon *YAP1* overexpression in both of two duplicate experiments, and for which the average increase in mRNA level in the two experiments was greater than threefold (50). Positions of the canonical *Yap1* binding sites upstream of the start codon, when present, and the average fold-increase in mRNA levels measured in the two experiments are indicated.

ORF	Distance of <i>Yap1</i> site from ATG	Gene	Description	Fold-increase
YNL331C			Putative aryl-alcohol reductase	12.9
YKL071W	162-222 (5 sites)	<i>YAP1</i>	Similarity to bacterial csgA protein	10.4
YML007W			Transcriptional activator involved in oxidative stress response	9.8
YFL056C	223, 242		Homology to aryl-alcohol dehydrogenases	9.0
YLL060C	98		Putative glutathione transferase	7.4
YOL165C	266		Putative aryl-alcohol dehydrogenase (NADP+)	7.0
YCR107W			Putative aryl-alcohol reductase	6.5
YML116W	409	<i>ATR1</i>	Aminotriazole and 4-nitroquinoline resistance protein	6.5
YBR008C	142, 167, 364		Homology to benomyl/methotrexate resistance protein	6.1
YCLX08C			Hypothetical protein	6.1
YJR155W			Putative aryl-alcohol dehydrogenase	6.0
YPL171C	148, 212	<i>OYE3</i>	NAPDH dehydrogenase (old yellow enzyme), isoform 3	5.8
YLR460C	167, 317		Homology to hypothetical proteins YCR102c and YNL134c	4.7
YKR076W	178		Homology to hypothetical protein YMR251w	4.5
YHR179W	327	<i>OYE2</i>	NAD(P)H oxidoreductase (old yellow enzyme), isoform 1	4.1
YML131W	507		Similarity to <i>A. thaliana</i> zeta-crystallin homolog	3.7
YOL126C		<i>MDH2</i>	Malate dehydrogenase	3.3

ing sites upstream of the others may reflect an ability of Yap1 to bind sites that differ from the canonical binding sites, perhaps in cooperation with other factors, or less likely, may represent an indirect effect of Yap1 overexpression, mediated by one or more intermediary factors. Yap1 sites were found only four times in the corresponding region of an arbitrary set of 30 genes that were not differentially regulated by Yap1.

Use of a DNA microarray to characterize the transcriptional consequences of mutations affecting the activity of regulatory molecules provides a simple and powerful approach to dissection and characterization of regulatory pathways and net-

works. This strategy also has an important practical application in drug screening. Mutations in specific genes encoding candidate drug targets can serve as surrogates for the ideal chemical inhibitor or modulator of their activity. DNA microarrays can be used to define the resulting signature pattern of alterations in gene expression, and then subsequently used in an assay to screen for compounds that reproduce the desired signature pattern.

DNA microarrays provide a simple and economical way to explore gene expression patterns on a genomic scale. The hurdles to extending this approach to any other organism are minor. The equipment

required for fabricating and using DNA microarrays (9) consists of components that were chosen for their modest cost and simplicity. It was feasible for a small group to accomplish the amplification of more than 6000 genes in about 4 months and, once the amplified gene sequences were in hand, only 2 days were required to print a set of 110 microarrays of 6400 elements each. Probe preparation, hybridization, and fluorescent imaging are also simple procedures. Even conceptually simple experiments, as we described here, can yield vast amounts of information. The value of the information from each experiment of this kind will progressively increase as more is learned about the functions of each gene and as additional experiments define the global changes in gene expression in diverse other natural processes and genetic perturbations. Perhaps the greatest challenge now is to develop efficient methods for organizing, distributing, interpreting, and extracting insights from the large volumes of data these experiments will provide.

REFERENCES AND NOTES

1. M. Schena, D. Shalon, R. W. Davis, P. O. Brown, *Science* 270, 467 (1995).
2. D. Shalon, S. J. Smith, P. O. Brown, *Genome Res.* 6, 639 (1996).
3. D. Lashkari, *Proc. Natl. Acad. Sci. U.S.A.*, in press.
4. J. DeRisi et al., *Nature Genet.* 14, 457 (1996).
5. D. J. Lockhart et al., *Nature Biotechnol.* 14, 1675 (1996).
6. M. Chee et al., *Science* 274, 610 (1996).
7. M. Johnston and M. Carlson, in *The Molecular Biology of the Yeast Saccharomyces: Gene Expression*, E. W. Jones, J. R. Pringle, J. R. Broach, Eds. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1992), p. 193.
8. Primers for each known or predicted protein coding sequence were supplied by Research Genetics. PCR was performed with the protocol supplied by Research Genetics, using genomic DNA from yeast strain S288C as a template. Each PCR product was verified by agarose gel electrophoresis and was deemed correct if the lane contained a single band of appropriate mobility. Failures were marked as such in the database. The overall success rate for a single-pass amplification of 6116 ORFs was ~94.5%.
9. Glass slides (Gold Seal) were cleaned for 2 hours in a solution of 2 N NaOH and 70% ethanol. After rinsing in distilled water, the slides were then treated with a 1:5 dilution of poly-L-lysine adhesive solution (Sigma) for 1 hour, and then dried for 5 min at 40°C in a vacuum oven. DNA samples from 100- μ l PCR reactions were purified by ethanol precipitation in 96-well microtiter plates. The resulting precipitates were resuspended in 3 \times standard saline citrate (SSC) and transferred to new plates for arraying. A custom-built arraying robot was used to print on a batch of 110 slides. Details of the design of the microarrayer are available at cmgm.stanford.edu/pbrown. After printing, the microarrays were rehydrated for 30 s in a humid chamber and then snap-dried for 2 s on a hot plate (100°C). The DNA was then ultraviolet (UV)-crosslinked to the surface by subjecting the slides to 60 mJ of energy (Stratagene Stratalinker). The rest of the poly-L-lysine surface was blocked by a 15-min incubation in a solution of 70 mM succinic anhydride dissolved in a solution consisting of 315 ml of 1-methyl-2-pyrrolidinone (Aldrich) and 35 ml of 1 M boric acid (pH 8.0). Directly after the blocking reac-

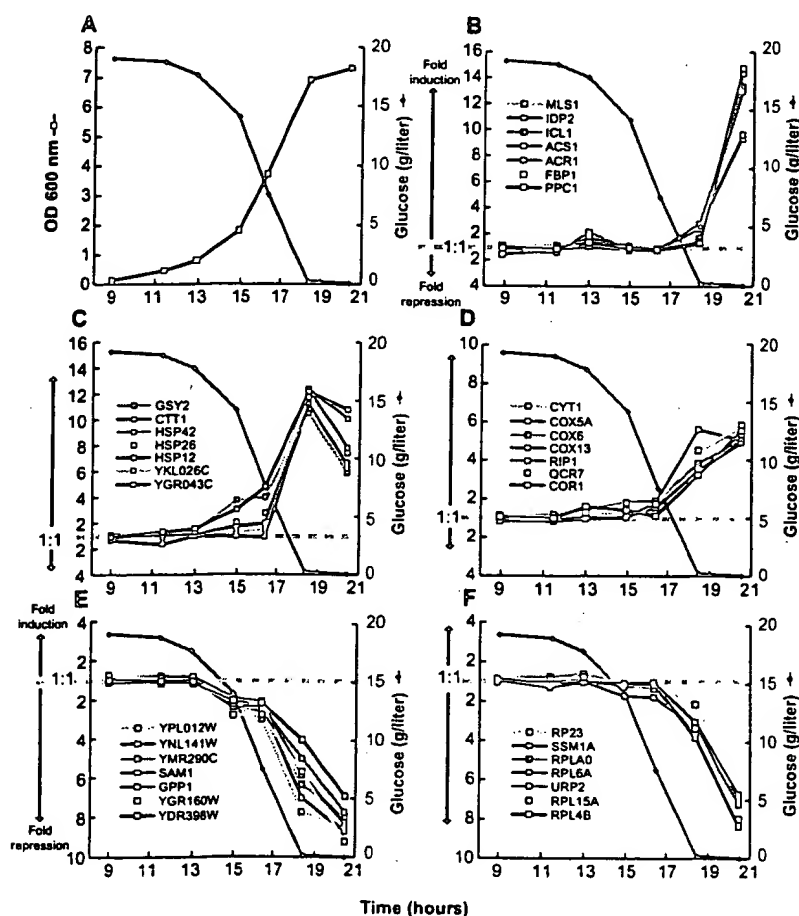


Fig. 5. Distinct temporal patterns of induction or repression help to group genes that share regulatory properties. (A) Temporal profile of the cell density, as measured by OD at 600 nm and glucose concentration in the media. (B) Seven genes exhibited a strong induction (greater than ninefold) only at the last timepoint (20.5 hours). With the exception of *IDP2*, each of these genes has a CSRE UAS. There were no additional genes observed to match this profile. (C) Seven members of a class of genes marked by early induction with a peak in mRNA levels at 18.5 hours. Each of these genes contain STRE motif repeats in their upstream promoter regions. (D) Cytochrome c oxidase and ubiquinol cytochrome c reductase genes. Marked by an induction coincident with the diauxic shift, each of these genes contains a consensus binding motif for the HAP2,3,4 protein complex. At least 17 genes shared a similar expression profile. (E) *SAM1*, *GPP1*, and several genes of unknown function are repressed before the diauxic shift, and continue to be repressed upon entry into stationary phase. (F) Ribosomal protein genes comprise a large class of genes that are repressed upon depletion of glucose. Each of the genes profiled here contains one or more RAP1-binding motifs upstream of its promoter. RAP1 is a transcriptional regulator of most ribosomal proteins.

- tion, the bound DNA was denatured by a 2-min incubation in distilled water at -95°C . The slides were then transferred into a bath of 100% ethanol at room temperature, rinsed, and then spun dry in a clinical centrifuge. Slides were stored in a closed box at room temperature until used.
10. YPD medium (8 liters), in a 10-liter fermentation vessel, was inoculated with 2 ml of a fresh overnight culture of yeast strain DBY7286 (MATa, ura3, GAL2). The fermentor was maintained at 30°C with constant agitation and aeration. The glucose content of the media was measured with a UV test kit (Boehringer Mannheim, catalog number 716251). Cell density was measured by OD at 600-nm wavelength. Aliquots of culture were rapidly withdrawn from the fermentation vessel by peristaltic pump, spun down at room temperature, and then flash frozen with liquid nitrogen. Frozen cells were stored at -80°C .
 11. Cy3-dUTP or Cy5-dUTP (Amersham) was incorporated during reverse transcription of 1.25 μg of polyadenylated [poly(A)⁺] RNA, primed by a dT(16) oligomer. This mixture was heated to 70°C for 10 min, and then transferred to ice. A premixed solution, consisting of 200 U Superscript II (Gibco), buffer, deoxyribonucleoside triphosphates, and fluorescent nucleotides, was added to the RNA. Nucleotides were used at these final concentrations: 500 μM for dATP, dCTP, and dGTP and 200 μM for dTTP. Cy3-dUTP and Cy5-dUTP were used at a final concentration of 100 μM . The reaction was then incubated at 42°C for 2 hours. Unincorporated fluorescent nucleotides were removed by first diluting the reaction mixture with of 470 μl of 10 mM Tris-HCl (pH 8.0)/1 mM EDTA and then subsequently concentrating the mix to ~ 5 μl , using Centricon-30 microconcentrators (Amicon).
 12. Purified, labeled cDNA was resuspended in 11 μl of $3.5\times$ SSC containing 10 μg poly(dA) and 0.3 μl of 10% SDS. Before hybridization, the solution was boiled for 2 min and then allowed to cool to room temperature. The solution was applied to the microarray under a cover slip, and the slide was placed in a custom hybridization chamber which was subsequently incubated for ~ 8 to 12 hours in a water bath at 62°C . Before scanning, slides were washed in $2\times$ SSC, 0.2% SDS for 5 min, and then $0.05\times$ SSC for 1 min. Slides were dried before scanning by centrifugation at 500 rpm in a Beckman CS-6R centrifuge.
 13. The complete data set is available on the Internet at cmgm.stanford.edu/pbrown/explore/index.html
 14. For 95% of all the genes analyzed, the mRNA levels measured in cells harvested at the first and second interval after inoculation differed by a factor of less than 1.5. The correlation coefficient for the comparison between mRNA levels measured for each gene in these two different mRNA samples was 0.98. When duplicate mRNA preparations from the same cell sample were compared in the same way, the correlation coefficient between the expression levels measured for the two samples by comparative hybridization was 0.99.
 15. The numbers and identities of known and putative genes, and their homologies to other genes, were gathered from the following public databases: Saccharomyces Genome Database (genome-www.stanford.edu), Yeast Protein Database (quest7.proteome.com), and Munich Information Centre for Protein Sequences (speedy.mips.biochem.mpg.de/mips/yeast/index.html).
 16. A. Scholer and H. J. Schuller, *Mol. Cell. Biol.* 14, 3613 (1994).
 17. S. Kratzer and H. J. Schuller, *Gene* 161, 75 (1995).
 18. R. J. Haselbeck and H. L. McAlister, *J. Biol. Chem.* 268, 12116 (1993).
 19. M. Fernandez, E. Fernandez, R. Rodicio, *Mol. Gen. Genet.* 242, 727 (1994).
 20. A. Hartig et al., *Nucleic Acids Res.* 20, 5677 (1992).
 21. P. M. Martinez et al., *EMBO J.* 15, 2227 (1996).
 22. J. C. Varela, U. M. Prætkelt, P. A. Meacock, R. J. Planta, W. H. Mager, *Mol. Cell. Biol.* 15, 6232 (1995).
 23. H. Ruis and C. Schuller, *Bioessays* 17, 959 (1995).
 24. J. L. Parrou, M. A. Teste, J. Francois, *Microbiology* 143, 1891 (1997).
 25. This expression profile was defined as having an induction of greater than 10-fold at 18.5 hours and less than 11-fold at 20.5 hours.
 26. S. L. Forsburg and L. Guarente, *Genes Dev.* 3, 1166 (1989).
 27. J. T. Olesen and L. Guarente, *ibid.* 4, 1714 (1990).
 28. M. Rosenkrantz, C. S. Kell, E. A. Pennell, L. J. Devenish, *Mol. Microbiol.* 13, 119 (1994).
 29. Single-letter abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr. The nucleotide codes are as follows: B-C, G, or T; N-G, A, T, or C; R-A or G; and Y-C or T.
 30. C. Fondrat and A. Kalogeropoulos, *Comput. Appl. Biosci.* 12, 363 (1996).
 31. D. Shore, *Trends Genet.* 10, 408 (1994).
 32. R. J. Planta and H. A. Raue, *ibid.* 4, 64 (1988).
 33. The degenerate consensus sequence YYCYRNNC-MNH was used to search for potential RAP1-binding sites. The exact consensus, as defined by (30), is WACAYCCRTACATYCW, with up to three differences allowed.
 34. S. F. Neuman, S. Bhattacharya, J. R. Broach, *Mol. Cell. Biol.* 15, 3187 (1995).
 35. P. Lesage, X. Yang, M. Carlson, *ibid.* 16, 1921 (1996).
 36. For example, we observed large inductions of the genes coding for *PCK1*, *FBP1* [Z. Yin et al., *Mol. Microbiol.* 20, 751 (1996)], the central glyoxylate cycle gene *ICL1* [A. Scholer and H. J. Schuller, *Curr. Genet.* 23, 375 (1993)], and the "aerobic" isoform of acetyl-CoA synthase, *ACS1* [M. A. van den Berg et al., *J. Biol. Chem.* 271, 28953 (1996)], with concomitant down-regulation of the glycolytic-specific genes *PFK1* and *PFK2* [P. A. Moore et al., *Mol. Cell. Biol.* 11, 5330 (1991)]. Other genes not directly involved in carbon metabolism but known to be induced upon nutrient limitation include genes encoding cytosolic catalase *CTT1* [P. H. Bissinger et al., *ibid.* 9, 1309 (1989)] and several genes encoding small heat-shock proteins, such as *HSP12*, *HSP26*, and *HSP42* [J. Farkas et al., *J. Biol. Chem.* 266, 15602 (1991); U. M. Prætkelt and P. A. Meacock, *Mol. Gen. Genet.* 223, 97 (1990); D. Wotton et al., *J. Biol. Chem.* 271, 2717 (1996)].
 37. The levels of induction we measured for genes that were expressed at very low levels in the uninduced state (notably, *FBP1* and *PCK1*) were generally lower than those previously reported. This discrepancy was likely due to the conservative background subtraction method we used, which generally resulted in overestimation of very low expression levels (46).
 38. Cross-hybridization of highly related sequences can also occasionally obscure changes in gene expression, an important concern where members of gene families are functionally specialized and differentially regulated. The major alcohol dehydrogenase genes, *ADH1* and *ADH2*, share 88% nucleotide identity. Reciprocal regulation of these genes is an important feature of the diauxic shift, but was not observed in this experiment, presumably because of cross-hybridization of the fluorescent cDNAs representing these two genes. Nevertheless, we were able to detect differential expression of closely related isoforms of other enzymes, such as *HXK1/HXK2* (77% identical) [P. Herrero et al., *Yeast* 11, 137 (1995)], *MLS1/DAL7* (73% identical) (20), and *PGM1/PGM2* (72% identical) [D. Oh, J. E. Hopper, *Mol. Cell. Biol.* 10, 1415 (1990)], in accord with previous studies. Use in the microarray of deliberately selected DNA sequences corresponding to the most divergent segments of homologous genes, in lieu of the complete gene sequences, should relieve this problem in many cases.
 39. F. E. Williams, U. Varanasi, R. J. Trumbly, *Mol. Cell. Biol.* 11, 3307 (1991).
 40. D. Tzamaras and K. Struhl, *Nature* 369, 758 (1994).
 41. Differences in mRNA levels between the *tup1Δ* and wild-type strain were measured in two independent experiments. The correlation coefficient between the complete sets of expression ratios measured in these duplicate experiments was 0.83. The concordance between the sets of genes that appeared to be induced was very high between the two experiments. When only the 355 genes that showed at least a twofold increase in mRNA in the *tup1Δ* strain in either of the duplicate experiments were compared, the correlation coefficient was 0.82.
 42. The *tup1Δ* mutation consists of an insertion of the LEU2 coding sequence, including a stop codon, between the ATG of *TUP1* and an Eco RI site 124 base pairs before the stop codon of the *TUP1* gene.
 43. L. R. Kowalski, K. Kondo, M. Inouye, *Mol. Microbiol.* 15, 341 (1995).
 44. M. Viswanathan, G. Muthukumar, Y. S. Cong, J. Lenard, *Gene* 148, 149 (1994).
 45. D. Hirata, K. Yano, T. Miyakawa, *Mol. Gen. Genet.* 242, 250 (1994).
 46. A. Gutierrez, L. Caramelo, A. Prieto, M. J. Martinez, A. T. Martinez, *Appl. Environ. Microbiol.* 60, 1783 (1994).
 47. A. Muheim et al., *Eur. J. Biochem.* 195, 369 (1991).
 48. J. A. Wemmie, M. S. Szczypka, D. J. Thiele, W. S. Moye-Rowley, *J. Biol. Chem.* 269, 32592 (1994).
 49. Microarrays were scanned using a custom-built scanning laser microscope built by S. Smith with software written by N. Ziv. Details concerning scanner design and construction are available at cmgm.stanford.edu/pbrown. Images were scanned at a resolution of 20 μm per pixel. A separate scan, using the appropriate excitation line, was done for each of the two fluorophores used. During the scanning process, the ratio between the signals in the two channels was calculated for several array elements containing total genomic DNA. To normalize the two channels with respect to overall intensity, we then adjusted photomultiplier and laser power settings such that the signal ratio at these elements was as close to 1.0 as possible. The combined images were analyzed with custom-written software. A bounding box, fitted to the size of the DNA spots in each quadrant, was placed over each array element. The average fluorescent intensity was calculated by summing the intensities of each pixel present in a bounding box, and then dividing by the total number of pixels. Local area background was calculated for each array element by determining the average fluorescent intensity for the lower 20% of pixel intensities. Although this method tends to underestimate the background, causing an underestimation of extreme ratios, it produces a very consistent and noise-tolerant approximation. Although the analog-to-digital board used for data collection possesses a wide dynamic range (12 bits), several signals were saturated (greater than the maximum signal intensity allowed) at the chosen settings. Therefore, extreme ratios at bright elements are generally underestimated. A signal was deemed significant if the average intensity after background subtraction was at least 2.5-fold higher than the standard deviation in the background measurements for all elements on the array.
 50. In addition to the 17 genes shown in Table 1, three additional genes were induced by an average of more than threefold in the duplicate experiments, but in one of the two experiments, the induction was less than twofold (range 1.6- to 1.9-fold).
 51. We thank H. Bennett, P. Spellman, J. Ravetto, M. Eisen, R. Pillai, B. Dunn, T. Ferea, and other members of the Brown lab for their assistance and helpful advice. We also thank S. Friend, D. Botstein, S. Smith, J. Hudson, and D. Dolginow for advice, support, and encouragement; K. Struhl and S. Chatterjee for the *Tup1* deletion strain; L. Fernandes for helpful advice on Yap1; and S. Klapholz and the reviewers for many helpful comments on the manuscript. Supported by a grant from the National Human Genome Research Institute (NHGRI) (HG00450), and by the Howard Hughes Medical Institute (HHMI). J.D.R. was supported by the HHMI and the NHGRI. V.R. was supported in part by an Institutional Training Grant in Genome Science (T32 HG00044) from the NHGRI. P.O.B. is an associate investigator of the HHMI.

5 September 1997; accepted 22 September 1997

The New York Times

ON THE WEB

October 2, 2003, Thursday

BUSINESS/FINANCIAL DESK

Human Genome Placed on Chip; Biotech Rivals Put It Up for Sale

By ANDREW POLLACK (NYT) 1030 words

The genome on a chip has arrived.

Melding high technology with biology, several companies are rushing to sell slivers of glass or nylon, some as small as postage stamps, packed with pieces of all 30,000 or so known human genes.

The new products will allow scientists to scan all genes in a human tissue sample at once, to determine which genes are active, a job that previously required two or more chips. The whole-genome chips will lower the cost and increase the speed of a widely used test that has transformed biomedical research in the last few years.

"It's sort of a milestone event, very similar to generating an integrated circuit of the genome," said Stephen P. A. Fodor, the chief executive of Affymetrix Inc., the leading seller of gene chips, which are also called microarrays.

Affymetrix, based in Santa Clara, Calif., is expected to announce today that it is accepting orders for its whole-genome chip.

The announcement seems timed to steal some thunder from the rival Agilent Technologies, which is based in nearby Palo Alto. Agilent is to be the host of an analyst meeting today and it plans to announce then that it has started shipping test versions of its whole-genome chip.

Applied Biosystems of Foster City, Calif., a unit of the Applera Corporation, started the race in July with an announcement that it would have a whole-genome chip out by the end of this year. NimbleGen Systems, a small company in Madison, Wis., announced a few days later that it had a genome on a chip that it was not selling but that it was using to run tests for customers.

Gene chips, which detect genes that are active, meaning they are being used to make a protein, have become essential tools. Scientists try to understand the genetic mechanisms of disease by seeing which genes are turned on in, say, a sick kidney or lung compared with those active in a healthy organ. Pharmaceutical companies look at gene activity patterns to try to predict the effects of drugs.

Scientists have found that tumors that look the same under the microscope can differ in terms of which genes are active. So studying gene patterns could become a way to discriminate between deadly and not-so-deadly tumors, or to predict which drug will work best for a particular patient.

Still, even some vendors conceded that the change from two chips to one is more symbolic than revolutionary.

"You can do just as good science with two chips, it costs you a little more," said Roland Green, the vice president for research and development at NimbleGen.

Some scientists questioned whether the chips really have all human genes, because the exact number and identities of all the genes is not known.

The advent of the genome on a chip is, however, evidence that biotechnology, to the extent that it uses electronics, is experiencing some of the rapid progress that has made semiconductors and computers continuously cheaper and smaller.

"One of the effects everyone is looking for in the genomics area is Moore's law -- more data, less money," said Doug Dolginow, an executive vice president at Gene Logic, which sells data from gene chip studies to pharmaceutical companies. "This is a step in that direction."

Moore's law states that the number of transistors on a semiconductor chip doubles every 18 months.

Affymetrix's gene chips are, in fact, made with the same techniques used to make semiconductor chips. In the mid-1990's, the company came out with a set of five chips covering what was then known of the human genome. After the human genome sequence was virtually completed in 2000, the company developed a two-chip set with all the known genes. Now it has the single chip, which some scientists say will be more convenient.

"We like to be able to look at all genes at one time to get a global view of what's going on," said John R. Walker, who runs gene chip operations at the Genomics Institute of the Novartis Research Foundation in San Diego.

Costs should also be lower. Gene chips have been so expensive that many academic scientists still make their own rather than buy them. Affymetrix said it would sell its whole-genome chips for \$300 to \$500 each, depending on volume, little more than half the price of the two-chip set. The other companies have not announced prices.

For Affymetrix, a successful whole-genome chip "is essential for them to maintain their dominance" of high-end microarrays, said Edward A. Tenthoff, an analyst at U.S. Bancorp Piper Jaffray. Affymetrix had total product sales in 2002 of about \$250 million, and a company spokesman said that human genome chips are its top-selling product.

Mr. Tenthoff, who recommends Affymetrix stock, said the company's sales growth rate had moderated as it faces tougher competition. Agilent, a spinoff of Hewlett-Packard that makes its gene chips by printing DNA components onto glass slides using ink jet printers, has gained share, he said. Applied Biosystems, the largest maker of genomics equipment over all, will be

entering the microarray segment of the business with its whole-genome chip, emphasizing the connection of that product to the others it offers, including the gene database developed by its sister company, Celera Genomics.

Jeffrey Trent, scientific director of the Translational Genomics Research Institute in Phoenix, said that while whole-genome chips are useful for medical discovery, the biggest growth of the market will be for chips that can be used by doctors to do diagnoses. And whole-genome chips are too cumbersome for that, he said. Rather, once scientists use the whole-genome chips to find particular genes that are associated with, say, tumor aggressiveness or drug effectiveness, he said, they will then make smaller and cheaper chips containing just those genes for use in diagnosis.



Agilent Technologies

About Agilent | Products & Services | Industries | International | Online Stores

[Worldwide Home](#) > [About Agilent](#) > [News@Agilent](#) > [Press Releases](#)

News@Agilent

Agilent Technologies ships whole human genome on single microarray to gene expression customers for evaluation

Company to introduce first commercial whole human microarray by end of year

PALO ALTO, Calif., Oct. 2, 2003

Press Releases

- ▶ [Community](#)
- ▶ [Corporate](#)
- ▶ [Electronic](#)
- ▶ [Life Sciences](#)
- ▶ [Chemical](#)

▶ [Archives](#)

Search Agilent

Quick Links

[Jump to page](#)

Agilent Technologies Inc. (NYSE: A) today announced it has shipped whole human-genome microarrays to customers for testing and evaluation. The whole genome microarray is based on Agilent's new double-density format, which can accommodate 44,000 features on a single 1" x 3" glass-slide microarray. The new platform enables drug-discovery and disease researchers to perform whole-genome screening at a lower cost and with higher reproducibility.

"This is an important step toward our release of the first whole human-genome microarray product, which is expected to be available for order before the end of the year," said Barney Saunders, vice president and general manager of Agilent's BioResearch Solutions Unit. "Customers have long wanted a one-sample, one-chip format with the increased sensitivity associated with 60-mer probes. The cost savings and high-quality performance make this product a compelling alternative for scientists who make their own microarrays."

Agilent's microarrays are based on the industry-standard 1" x 3" (25mm x 75mm) format, which is compatible with most commercial microarray scanners. All Agilent commercial microarrays are developed using content from public databases and proprietary sources, with full sequence and annotation information made available to customers. Gene sequences for probes are developed using algorithms and then validated empirically through iterative wet-lab testing procedures. The result is a microarray comprised of functionally validated probes, with the most up-to-date and comprehensive genome information commercially available.

Advantages of the double-density format include:

- Lower cost. Not only is one microarray less expensive than two, it requires fewer reagents and reduces instrumentation demands.
- Streamlined workflow. Researchers need prepare and process only one microarray instead of two. This also results in fewer steps in the subsequent data analysis.
- Greater reproducibility. Use of a single microarray further reduces unnecessary variability in experimental conditions.
- Smaller sample use. A smaller quantity of sample material is required to perform an experiment.

Availability

Agilent's Whole Human Genome Microarray is expected to be available for order by the end of the year.

About Agilent Technologies

Agilent Technologies Inc. (NYSE: A) is a global technology leader in communications, electronics, life sciences and chemical analysis. The company's 30,000 employees serve customers in more than 110 countries. Agilent had net revenue of \$6 billion in fiscal year 2002. Information about Agilent is available

on the Web at www.agilent.com.

Forward-Looking Statements

This news release contains forward-looking statements (including, without limitation, statements relating to Agilent's expectation that its whole-genome microarray platform will be available for order before the end of 2003) that involve risks and uncertainties that could cause results to differ materially from management's current expectations. These and other risks are detailed in the company's filings with the Securities and Exchange Commission, including its Annual Report on Form 10-K for the year ended Oct. 31, 2002, its Quarterly Report on Form 10-Q for the quarter ended July 31, 2003 and its Current Report on Form 8-K filed Aug. 18, 2003. The company assumes no obligation to update the information in this press release.

###

Contact:

Christina Maehr
+1 408 553 7205
christina_maehr@agilent.com

To send feedback about this site: [Contact Webmaster](#)

© Agilent 2000-2003

[Terms of Use](#)

[Privacy](#)

Today's News

Affymetrix Announces Commercial Launch of Single Array for Human Genome Expression Analysis**AFFYMETRIX GENECHIP(R) BRAND HUMAN GENOME U133 PLUS 2.0 ARRAY**

Affymetrix GeneChip(R) Brand Human Genome U133 Plus 2.0 Array.
(PRNewsFoto)[AS]
SANTA CLARA, CA USA 10/02/2003

[Website](#)

More Than 1 Million Probes Analyze Expression Levels of Nearly 50,000 RNA Transcripts and Variants on a Single Array the Size of a Thumbnail

SANTA CLARA, Calif., Oct. 2 /PRNewswire/ -- Affymetrix, Inc., (Nasdaq: AFFX) announced today that it is taking orders for its new GeneChip(R) brand Human Genome U133 Plus 2.0 Array, offering researchers the protein-coding content of the human genome on a single commercially available catalog microarray. The HG-U133 Plus 2.0 Array analyzes the expression level of nearly 50,000 RNA transcripts and variants with 22 different probes per transcript, providing superior data quality unmatched by technologies using a single probe per transcript.

(Photo: <http://www.newscom.com/cgi-bin/prnh/20031002/SFTH021>)

"With about 1.3 million probes on a chip the size of a human thumbnail, the Human Plus Array represents a leap in array technology data capacity, and further demonstrates the unique power and potential of our technology to explore vast areas of the genome," said Trevor J. Nicholls, Ph.D., Chief Commercial Officer. "Multiple independent measurements for each transcript ensure that our data quality remains the industry standard, even as our data capacity increases dramatically."

The HG-U133 Plus 2.0 Array, which will ship in October, combines the content of the previous HG-U133 two-array set with nearly 10,000 new probe sets representing about 6,500 new genes, for a total of nearly 50,000 RNA transcripts and variants. This new information, verified against the latest version of the publicly available genome map, provides researchers the most comprehensive and up-to-date genome-wide gene expression analysis. The probe design strategy of the HG-U133 Plus 2.0 Array is identical to the previous HG-U133 Set, providing very strong data concordance between the two products. With more than double the data capacity of the previous-generation Affymetrix human product, the HG-U133 Plus 2.0 Array can significantly cut processing and analysis time for scientists in the lab, freeing up valuable resources and accelerating research.

The HG-U133 Plus 2.0 Array sets a new standard for the number of genes and transcripts on any commercially available single array for human gene

expression analysis, while maintaining Affymetrix' unrivaled data quality. The HG-U133 Plus 2.0 Array uses 22 independent measures to detect the hybridization of each transcript on the array, 1.3 million data points in all, more than 30 times that of any other microarray technology. Using multiple, independent measurements provides optimal sensitivity and specificity, and the most accurate, consistent and statistically significant results possible.

"More data points produce more reliable results and ultimately, enable better science," said Nicholls. "Our powerful probe set strategy gives our customers the assurance that their array results actually reflect what's in their sample."

Affymetrix is also launching an updated 11-micron version of its popular 18-micron HG-U133A Array called the GeneChip HG-U133A 2.0 Array. The reduced feature size on this new design means researchers can use smaller sample volumes than on the previous 18-micron array without compromising performance. This new array represents over 20,000 transcripts that can be used to explore human biology and disease processes. All probe sets represented on the original GeneChip HG-U133A Array are identically replicated on the GeneChip HG-U133A 2.0 Array.

More information on the design of the HG-U133 Plus 2.0 Array and the HG-U133A 2.0 Array may be found on the Affymetrix website at <http://www.affymetrix.com>.

Affymetrix will be presenting further information on this and other products at the BioTechnica trade show in Hanover, Germany on Oct. 7-9, 2003. The Company will also hold a press conference on Oct. 7, from 11 a.m. to 12 p.m. at the show regarding the new Human Genome U133 Plus 2.0 Array. If you would like to attend this press conference, please contact Caroline Stupnicka at c.stupnicka@northbankcommunications.com.

About Affymetrix:

Affymetrix is a pioneer in creating breakthrough tools that are driving the genomic revolution. By applying the principles of semiconductor technology to the life sciences, Affymetrix develops and commercializes systems that enable scientists to improve the quality of life. The Company's customers include pharmaceutical, biotechnology, agrichemical, diagnostics and consumer products companies as well as academic, government and other non-profit research institutes. Affymetrix offers an expanding portfolio of integrated products and services, including its integrated GeneChip platform, to address growing markets focused on understanding the relationship between genes and human health. Additional information on Affymetrix can be found at <http://www.affymetrix.com>.

All statements in this press release that are not historical are "forward-looking statements" within the meaning of Section 21E of the Securities Exchange Act as amended, including statements regarding Affymetrix' "expectations," "beliefs," "hopes," "intentions," "strategies" or the like. Such statements are subject to risks and uncertainties that could cause actual results to differ materially for Affymetrix from those projected, including, but not limited to risks of the Company's ability to achieve and sustain higher levels of revenue, higher gross margins, reduced operating expenses, uncertainties relating to technological approaches, manufacturing, product development, market acceptance (including uncertainties relating to product development and market acceptance of the GeneChip HG-U133 Human Plus 2.0 Array and the HG-U133A 2.0), personnel retention, uncertainties related to cost and pricing of Affymetrix products, dependence on collaborative partners, uncertainties relating to sole source suppliers, uncertainties relating to FDA and other regulatory approvals, competition, risks relating to intellectual property of others and the uncertainties of patent protection and litigation. These and other risk factors are discussed in Affymetrix' Form 10-K for the

year ended December 31, 2002 and other SEC reports, including its Quarterly Reports on Form 10-Q for subsequent quarterly periods. Affymetrix expressly disclaims any obligation or undertaking to release publicly any updates or revisions to any forward-looking statements contained herein to reflect any change in Affymetrix' expectations with regard thereto or any change in events, conditions, or circumstances on which any such statements are based.

NOTE: Affymetrix, the Affymetrix logo, and GeneChip and are registered trademarks owned or used by Affymetrix, Inc.

SOURCE Affymetrix, Inc.

Web Site: <http://www.affymetrix.com>

Photo Notes: NewsCom:

<http://www.newscom.com/cgi-bin/prnh/20031002/SFTH021> AP Archive:

<http://photoarchive.ap.org> PRN Photo Desk,

photodesk@prnewswire.com

Issuers of news releases and not PR Newswire are solely responsible for the accuracy of the content.

More news from PR Newswire...

Copyright © 1996-2002 PR Newswire Association LLC. All Rights Reserved.
A United Business Media company.

Macroresults through Microarrays

John C. Rockett, Reproductive Toxicology Division (MD-72), National Health and Environmental Effects Research Laboratory, Office of Research and Development, US Environmental Protection Agency, Research Triangle Park, 2525 East Highway 54, Durham, NC 27711, USA;
tel: +1 919 541 2071, fax: +1 919 541 4017, e-mail: rockett.john@epa.gov

The third enactment of Cambridge Healthtech Institute's *Macroresults through Microarrays* meeting was held in Boston (MA, USA) from 29 April–1 May 2002. The subtheme of this year's meeting was 'advancing drug discovery', a widely touted application for array technology.

The evolution of microarrays

If you were asked 'Who first conceived of the idea of microarrays', who would come to mind? Mark Schena perhaps, first author of the seminal 1995 paper on cDNA arrays [1]? Maybe Pat Brown, Schena's then supervisor? Or perhaps Stephen Fodor, the primary driver behind Affymetrix's (<http://www.affymetrix.com>) oligonucleotide-based platform [2]. Brits might even chant the name of Ed Southern [3]. Well, according to Roger Ekins (University College London Medical School; <http://www.ucl.ac.uk/medicine/>) all these answers would be wrong. It was in fact Ekins and his colleagues who first conceived of and patented 'a new generation of ultrasensitive, miniaturized assays for protein and DNA–RNA measurement based on the use of microarrays' in the mid 1980s [4]. The concept and potential of array technology was more fully described in a later publication, in which Ekins *et al.* [5] concluded that antibody microspots of $\sim 50 \mu\text{m}^2$ could be achieved, and that as many as 2 million different immunoassays could, in principle, be accommodated on a surface area of 1 cm^2 .

Technological innovation

In practice, it took a different biological molecule (DNA), a different research

group, and a leap into microfabrication technology to even begin approaching these kinds of densities [Affymetrix patent 6045996 talks of one million spots cm^{-2}]. Of course, advancing technology is one of the driving engines behind the genomics juggernaut, and we are already seeing '4th generation' machines for fabricating DNA chips. If the company representatives at this meeting are to be believed (and their cases seemed strong), spotting is out, and *in situ* fabrication of oligonucleotide-based 'iterative custom arrays' is in. Whether you go with the Combimatrix's (<http://www.combimatrix.com>) electrochemically directed synthesis and detection system, febit's (<http://www.febeit.com>) Geniom® technology, or Nimblegen's (<http://www.nimblegen.com>) Maskless Array Synthesizer technology is a matter of personal choice. However, each of these machines provides the flexibility to design variable length oligonucleotide probes from sequences inputted by the user, and then perform *in situ* synthesis of an array. Each system also boasts unique advantages. For example, Combimatrix's biological array processor is a semiconductor coated with a 3D layer of porous material in which DNA, RNA, peptides or small molecules can be synthesized or immobilized within discrete test sites, while febit's Geniom One® is a fully integrated gene-expression analysis system with minimal user hands-on time – the probe sequences are programmed, the RNA samples inserted, and the gene expression data is pumped out a few hours later.

Cell- and tissue-based arrays

Array technology is in most people's minds firmly linked with gene-expression profiling. Fewer are aware that cell- and tissue-based arrays have been developed, and how they can provide a vital extra dimension to research. In support of this, Barry Bochner gave an update on the cell-based array system that Biolog (<http://www.biolog.com>) has produced for simultaneously measuring the effects of one gene in the cell under thousands of growth conditions (see [6] for further details). David Walt (Tufts University; <http://www.tufts.edu/>) is developing single live cell arrays using optical imaging fiber (OIF) technology. An array of microwells is fabricated on the face of an OIF at densities of up to 10 million wells cm^{-2} . Cells are then added to the wells and disperse at an average of one cell per well. Physiological and genetic responses of each cell are measured via fluorescence produced by reporter genes (e.g. *lacZ*, *gfp*). Assays performed so far include yeast live or dead cell assay, microenvironment pH and O_2 measurements, promoter responses using the *lacZ* and *phoA* reporter genes, and protein–protein interactions using the yeast two-hybrid system. The main advantage of this system is that the cells remain alive during the assay, which means a real-time timecourse can be performed and/or the array passed from sample to sample. This would be useful in, for example, the scanning of a combinatorial drug library for specific physiological effects.

Tissue arrays are a useful complementary technology to DNA arrays because they can be used to help validate and

understand the biological and medical significance of gene changes discovered using standard DNA arrays. For example, an array of tumor tissues can be screened for the protein (using immunohistochemistry), message (using *in situ* hybridization) and copy number (using comparative genomic hybridization) of a gene of interest, to determine if expression of the gene (or lack thereof) is related in any way to survival. They can also be used to predict the probability of clinical failure of lead compounds as a result of toxicity by evaluating the distribution of the drug targets in normal tissue. Spyro Mousses and his co-workers at the National Human Genome Research Institute (<http://www.nhgri.nih.gov/index.html>) have built such arrays, including a multi-tumor array (~5000 specimens, and sections from 36 normal and 800 metastatic tissues) and a normal tissue array (76 tissue and 332 cell types).

The problem with proteins

It has been said that genomics tells us what might happen, transcriptomics indicates what should happen, and proteomics shows what is happening. The impact of functional proteomics on pharmaceutical R&D is rapidly increasing, and protein arrays are being used increasingly in both basic and applied research. Their use lies not only in comparative protein expression and interaction profiling, but also in diagnostics and drug discovery. However, an increasing number of researchers have found that protein arrays, like their cousins the DNA arrays, present several practical obstacles relating to their production and use. For example, in using *Escherichia coli* to produce recombinant eukaryotic proteins from a single expression vector, multiple protein products are often produced, suggesting mixes of truncated or otherwise altered proteins. There is also the obvious concern that the proteins might not be modified in a similar manner to

eukaryotic systems. Also, an optimal method for depositing and binding proteins to the selected substrate is yet to be determined, as is the best way to ensure that they are bound in a correctly folded, active conformation.

Several companies have been addressing these problems. Prolinx (<http://www.prolinxinc.com>) is one such company, and Karin Hughes described their Versalinx™ chemistry for producing protein, peptide and small-molecule arrays. Versalinx™ uses solution-phase conjugation followed by immobilization, resulting in functional orientation of proteins and peptides on the substrate surface. It also offers the valuable additional benefit of exhibiting low non-specific binding. Sense Proteomic (<http://www.senseproteomic.com>) is also among those addressing these problems to develop robust protein arrays for drug discovery and clinical applications and has developed functional protein array formats based on specific disease tissues. Subtractive hybridization is used to identify genes with altered expression in breast tumor and cystic fibrosis compared to normal tissue. A high throughput cloning strategy (COVET™) is then used to produce libraries of genes that are tagged, cloned, expressed, purified and finally immobilized on glass slides. Initial validation studies have shown that the vast majority of the immobilized proteins do indeed retain biological function.

Stefan Schmidt and his company (GPC Biotech; <http://www.gpcbiotech.de>) have moved past the platform development stage and, with their focus firmly on drug discovery, are currently developing kinase-profiling arrays. Kinases are important targets for pharmaceutical drug discovery and therapy, and GPC's aim is to simultaneously detect multiple kinases, obtain activity profiles for different cell types, or analyze the ability of drug candidates to inhibit kinase activity. To do this, recombinant kinase substrates are immobilized on

membranes, incubated with purified kinase, and the substrates measured for the degree of phosphorylation.

Summary

Meetings like this, packed with exciting discoveries and intriguing and interesting innovation, heavily emphasize the pace at which biotechnology is advancing, to the extent that the number of options for genomic and proteomic researchers can become overwhelming. Although data analysis is perhaps the greatest current concern for array users, an increasing challenge will be to determine the approaches and technology that really work, and to do it in a timely manner.

References

- 1 Schena, M. *et al.* (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270, 467-470
- 2 Fodor, S.P. *et al.* (1991) Light-directed, spatially addressable parallel chemical synthesis. *Science* 251, 767-773
- 3 Southern, E.M. *et al.* (1992) Analyzing and comparing nucleic acid sequences by hybridization to arrays of oligonucleotides: evaluation using experimental models. *Genomics* 13, 1008-1017
- 4 Ekins, R.P. (1987) US Patent Application 8 803 000
- 5 Ekins, R. *et al.* (1989) High specific activity chemiluminescent and fluorescent markers: their potential application to high sensitivity and 'multi-analyte' immunoassays. *J. Biol. Chem.* 264, 59-78
- 6 Rockett, J.C. (2002) Chip, chip, array! Three chips for post-genomic research. *Drug Discov. Today* 7, 458-459

Acknowledgements

I would like to thank Mary Ann Brown (Cambridge Healthtech Institute) and David Dix (US EPA) for critical review of this manuscript prior to submission. This document has been reviewed in accordance with US Environmental Protection Agency policy and approved for publication. Mention of companies, trade names or products does not signify endorsement of such by the EPA.

N. Leigh Anderson
Ricardo Esquer-Blasco
Jean-Paul Hofmann
Norman G. Anderson

Large Scale Biology Corporation,
Rockville, MD

A two-dimensional gel database of rat liver proteins useful in gene regulation and drug effects studies

A standard two-dimensional (2-D) protein map of Fischer 344 rat liver (F344MST3) is presented, with a tabular listing of more than 1200 protein species. Sodium dodecyl sulfate (SDS) molecular mass and isoelectric point have been established, based on positions of numerous internal standards. This map has been used to connect and compare hundreds of 2-D gels of rat liver samples from a variety of studies, and forms the nucleus of an expanding database describing rat liver proteins and their regulation by various drugs and toxic agents. An example of such a study, involving regulation of cholesterol synthesis by cholesterol-lowering drugs and a high-cholesterol diet, is presented. Since the map has been obtained with a widely used and highly reproducible 2-D gel system (the Iso-Dalt® system), it can be directly related to an expanding body of work in other laboratories.

Contents

1 Introduction.....	907
2 Material and methods	908
2.1 Sample preparation.....	908
2.2 Two-dimensional electrophoresis	909
2.3 Staining	909
2.4 Positional standardization	909
2.5 Computer analysis.....	909
2.6 Graphical data output	910
2.7 Experiment LSBC04	910
3 Results and discussion.....	910
3.1 The rat liver protein 2-D map.....	910
3.2 Carbamylated charge standards computed pI's and molecular mass standardization	911
3.3 An example of rat liver gene regulation: Chol- esterol metabolism	911
3.3.1 MSN 413 (putative cytosolic HMG-CoA synthase) and sets of spots regulated co- ordinately or inversely	911
3.3.2 MSN 235 and coregulated spots.....	912
3.3.3 An example of an anti-synergistic effect	912
3.3.4 Complexity of the cholesterol synthesis pathway	912
4 Conclusions	912
5 References	912
6 Addendum 1: Figures 1-13.....	914
7 Addendum 2: Tables 1-4.....	923
Table 1. Master table of proteins in rat liver data- base	923
Table 2. Table of some identified proteins	928
Table 3. Computed pI's of two sets of carbamylated protein standards: rabbit muscle CPK and human Hb.....	929
Table 4. Computed pI's of some known proteins re- lated to measured CPK pI's.....	930

1 Introduction

High-resolution two-dimensional electrophoresis of proteins, introduced in 1975 by O'Farrell and others [1-4], has been used over the ensuing 16 years to examine a wide variety of biological systems, the results appearing in more than 5000 published papers. With the advent of computerized systems for analyzing two-dimensional (2-D) gel images and constructing spot databases, it is also possible to plan and assemble integrated bodies of information describing the appearance and regulation of thousands of protein gene products [5, 6]. Creating such databases involves amassing and organizing quantitative data from thousands of 2-D gels, and requires a substantial commitment in technology and resources.

Given the long-term effort required to develop a protein database, the choice of a biological system takes on considerable importance. While *in vitro* systems are ideal for answering many experimental questions, especially in cancer research and genetics, our experience with cell cultures and tissue samples suggests that some *in vivo* approaches could have major advantages. In particular, we have noticed that liver tissue samples from rats and mice appear to show greater quantitative reproducibility (in terms of individual protein expression) than replicate cell cultures. This is perhaps a natural result of the homeostasis maintained in a complete animal vs. the well-known variability of cell cultures, the latter due principally to differences in reagents (e.g., fetal bovine serum), conditions (e.g., pH) and genetic "evolution" of cell lines while in culture. It is also more difficult to generate adequate amounts of protein from cell culture systems (particularly with attached cells), forcing the investigator to resort to radioisotope-based or silver-based stain-detection methods. While these methods are more sensitive (sometimes much more sensitive) than the Coomassie Brilliant Blue (CBB) stain typically used for protein detection in "large" protein samples, they are generally more variable, more labor-intensive and, in the case of radiographic methods, may generate highly "noisy" images, due to the properties of the films used. By contrast, large protein samples can easily be prepared from liver using urea/Nonidet P-40 (NP-40) solubilization and stained with CBB, which has the advantage of being easily reproducible [8]. Finally, there remains the question of the "truthfulness" of many *in vitro* systems as compared to their *in vivo* analogs; how great are the changes caused by the introduction into a cul-

Correspondence: Dr. N. Leigh Anderson, Large Scale Biology Corporation, 9620 Medical Center Drive, Rockville, MD 20850, USA

Abbreviations: CBB, Coomassie Brilliant Blue; CPK, creatine phosphokinase; 2-D, two-dimensional; IEF, isoelectric focusing; MSN, master spot number; NP-40, Nonidet P-40; SDS, sodium dodecyl sulfate

ture and the associated shift to strong selection for growth, and how do these affect experimental outcomes? Hence the apparent advantages of *in vitro* systems, in terms of experimental manipulation, may be counterbalanced by other factors relating to 2-D data quality.

There is a second important class of reasons for exploring the use of an *in vivo* biological system such as the liver. Historically, there have been two broad approaches to the mechanistic dissection of biochemical processes in intact cellular systems: genetics (a search for informative mutants) and the use of chemical agents (drugs and chemical toxins). Both approaches help us to understand complex systems by disrupting some specific functional element and showing us the result. With the development of techniques for genetic manipulation and cloning, the genetic approach can be effectively applied either *in vitro* or *in vivo*, although the *in vitro* route is usually quicker. The chemical approach can also be applied to either sort of biological system; here, however, the bulk of consistently acquired information is in experimental animals (rats and mice). While most biologists know a short list of compounds having specific, experimentally useful effects (e.g., inhibitors of protein synthesis, ionophores, polymerase inhibitors, channel blockers, nucleotide analogs, and compounds affecting polymerization of cytoskeletal proteins), there is a much larger number of interesting chemically-induced effects, most of them characterized by toxicologists and pharmacologists in rodent systems. Just as a thorough genetic analysis would involve saturating a genome with mutations, it is possible to imagine a saturating number of drugs, the analysis of whose actions would reveal the complete biochemistry of the cell. While organized drug discovery efforts usually target specific desired effects, the nature of the process, with its dependence on screening large numbers of compounds, necessarily produces many unanticipated effects. It is therefore reasonable to suppose that the required broad range of compounds necessary to achieve "biochemical saturation" may be forthcoming; in fact, it may already exist among the hundreds of thousands of compounds that failed to qualify as drugs.

Among organs, the liver is an obvious choice for the study of chemical effects because of its well-known plasticity and responsiveness. The brain appears to be quite plastic (e.g. [7]), but it is a complicated mixture of cell types requiring skillful dissection for most experiments. The kidney, while quite responsive, also presents a potentially confounding mixture of cell types. The liver, by contrast, is made up of one predominant cell type which is easy to solubilize: the hepatocyte, representing more than 95% of its mass. Most importantly, the liver performs many homeostatic functions that require rapid modulation of gene expression. It appears that most chemical agents tested affect gene expression in the liver at some dosage (N. Leigh Anderson, unpublished observations), an interesting contrast to our earlier work with lymphocytes, for example, which seem to be much less responsive. Such results conform to the expectation that cells with a homeostatic, physiological role should be more plastic than cells differentiated for a purpose dependent on the action of a limited number of specific genes.

The liver also allows the parallels between *in vitro* and *in vivo* systems to be examined in detail. Significant progress

has been made in the development of mouse, rat and human hepatocyte culture systems, as well as in precision-cut tissue slices. Using such an array of techniques, it is possible to assemble a matrix of mammalian systems including mouse and rat *in vivo* on one level and mouse, rat and human *in vitro* on a second level, and to compare effects between species and between systems. This approach allows us to draw informed conclusions regarding the biochemical "universality" of biological responses among the mammals, and to offer some insight into the validity of *in vitro* approaches for toxicological screening. We believe this data will be necessary if *in vitro* alternatives are to achieve wide usage in government-mandated safety testing of drugs, consumer products and industrial and agricultural chemicals.

A number of interesting studies have been published using 2-D mapping to examine effects in the rodent liver. A number of investigators have made use of the technique to screen for existing genetic variants [8-11] or induced mutations [12-14], mainly in the mouse. This work builds on the wealth of genetic information available on the mouse and its established position as a mammalian mutation-detection system. While some studies of chemical effects have been undertaken in the mouse [15-17], most have used the rat [18-23]. The examination of the cytochrome p-450 system, in particular, has been carried out almost exclusively on the rat [24, 25].

These considerations lead us to conclude that rodent liver offers the best opportunity to systematically examine an array of gene regulation systems, and ultimately to build a predictive model of large-scale mammalian gene control. The basic underlying foundation of such a project is a reliable, reproducible master 2-D pattern of liver, to which ongoing experimental results can be referred. In this paper, we report such a master pattern for the acidic and neutral proteins of rat liver (pattern F344MST3). In future, this master will be supplemented by maps of basic proteins, and analogous maps of mouse and human liver.

2 Materials and methods

2.1 Sample preparation

Liver is an ideal sample material for most biochemical studies, including 2-D analysis. A sample is taken of approximately 0.5 g of tissue from the apical end of the left lobe of the liver. Solubilization is effected as rapidly as practical; a delay of 5-15 min appears to cause no major alteration in liver protein composition if the liver pieces are kept cold (e.g., on ice) in the interim. In the solubilization process, the liver sample is weighed, placed in a glass homogenizer (e.g., 15 mL Wheaton); 8 volumes of solubilizing solution*

* The solubilizing solution is composed of 2% NP-40 (Sigma), 9 M urea (analytical grade, e.g., BDH or Bio-Rad), 0.5% dithiothreitol (DTT; Sigma) and 2% carrier ampholytes (pH 9-11 LKB; these come as a 20% stock solution, so 2% final concentration is achieved by making the final solution 10% 9-11 Ampholine by volume). A large batch of solubilizer (several hundred mL) is made and stored frozen at -80°C in aliquots sufficient to provide enough for one day's estimated sample preparation requirement. The solution is never allowed to become warmer than room temperature at any stage during preparation or thawing for use, since heating of concentrated urea solutions can produce contaminants that covalently modify proteins producing artifactual charge shifts. Once thawed, any unused solubilizer is discarded.

is added (i.e., 4 mL per 0.5 g tissue) and the mixture is homogenized using first the loose- and then the tight-fitting glass pestle. This takes approximately 5 strokes with each pestle and is carried out at room temperature because urea would crystallize out in the cold. Once the liver sample is thoroughly homogenized in the solubilizer, it is assumed that all the proteins are denatured (by the chaotropic effect of the urea and NP-40 detergent) and the enzymes inactivated by the high pH (~9.5). Therefore these samples may be kept at room temperature until they can be centrifuged or frozen as a group (within several hours of preparation). The samples are centrifuged for 6×10^6 g min (e.g., 500 000 \times g for 12 min using a Beckman TL-100 centrifuge). The centrifuge rotor is maintained at just below room temperature (e.g., 15–20°C), but not too cold, so as to prevent the precipitation of urea. The centrifuge of choice is a Beckman TL-100 because of the sample tube sizes available, but any ultracentrifuge accepting smallish tubes will suffice. When an appropriate centrifuge is not available near the site of sample preparation, samples can be frozen at –80°C and thawed prior to centrifugation and collection of supernatants. Each supernatant is carefully removed following centrifugation and aliquoted into at least 4 clean tubes for storage. This is done by transferring all the supernatant to one clean tube, mixing this gently (to assure homogeneous composition) and then dividing it into 4 aliquots. The aliquots are frozen immediately at –80°C. These multiple aliquots can provide insurance against a failed run or a freezer breakdown.

2.2 Two-dimensional electrophoresis

Sample proteins are resolved by 2-D electrophoresis using the 20 \times 25 cm Iso-Dalt® 2-D gel system ([26–29]; produced by LSB and by Hoefer Scientific Instruments, San Francisco) operating with 20 gels per batch. All first-dimensional isoelectric focusing (IEF) gels are prepared using the same single standardized batch of carrier ampholytes (BDH 4–8A in the present case, selected by LSB's batch-testing program for rat and mouse database work**). A 10 μ L sample of solubilized liver protein is applied to each gel, and the gels are run for 33 000 to 34 500 volt-hours using a progressively increasing voltage protocol implemented by a programmable high-voltage power supply. An Angeliq™ computer-controlled gradient-casting system (produced by LSB) is used to prepare second-dimensional sodium dodecyl sulfate (SDS) polyacrylamide gradient slab gels in which the top 5% of the gel is 11%T acrylamide, and the lower 95% of the gel varies linearly from 11% to 18%T.

This system has recently been modified so as to employ a commercially available 30.8%T acrylamide/*N,N*-methylenebisacrylamide prepared solution (thus avoiding the handling of the solid acrylamide monomer) and three additional stock solutions: buffer (made from Sigma pre-set Tris), persulfate and *N,N,N',N'*-tetramethylethylenediamine (TEMED). Each gel is identified by a computer-printed filter paper label polymerized into the lower left corner of the gel. First-dimensional IEF tube gels are loaded

directly (as extruded) onto the slab gels without equilibration, and held in place by polyester fabric wedges (Wedgies™, produced by LSB) to avoid the use of hot agarose. Second-dimensional slab gels are run overnight, in groups of 20, in cooled DALT tanks (10°C) with buffer circulation. All run parameters, reagent source and lot information, and notations of deviation from expected results are entered by the technician responsible on a detailed, multi-page record of the experiment.

2.3 Staining

Following SDS-electrophoresis, slab gels are stained for protein using a colloidal Coomassie Blue G-250 procedure in covered plastic boxes, with 10 gels (totalling approximately 1 L of gel) per box. This procedure (based on the work of Neuhoff [30, 31]) involves fixation in 1.5 L of 50% ethanol and 2% phosphoric acid for 2 h, three 30 min washes, each in 2 L of cold tap water, and transfer to 1.5 L of 34% methanol, 17% ammonium sulfate and 2% phosphoric acid for 1 h, followed by the addition of a gram of powdered Coomassie Blue G-250 stain. Staining requires approximately 4 days to reach equilibrium intensity, whereupon gels are transferred to cool tap water and their surfaces rinsed to remove any particulate stain prior to scanning. Gels may be kept for several months in water with added sodium azide. The water washes remove ethanol that would dissolve the stain (and render the system noncolloidal, with high backgrounds). The concentrated ammonium sulfate and methanol solution is diluted by equilibration with the water volume of the gels to automatically achieve the correct final concentrations for colloidal staining. Practical advantages of this staining approach can be summarized as follows: (i) the low, flat background makes computer evaluation of small spots (max OD < 0.02) possible, especially when using laser densitometry; (ii) up to 1500 spots can be reliably detected on many gels (e.g., rat liver) at loadings low enough to preserve excellent resolution; and (iii) reproducibility appears to be very good: at least several hundred spots have coefficients of reproducibility less than 15%. This value is at least as good as previous CBB methods, and significantly better than many silver stain systems.

2.4 Positional standardization

The carbamylated rabbit muscle creatine phosphokinase (CPK) standards [32] are purchased from Pharmacia and BDH. Amino acid compositions, and numbers of residues present in proteins used for internal standardization, are taken from the Protein Identification Resource (PIR) sequence database [33].

2.5 Computer analysis

Stained slab gels are digitized in red light at 134 micron resolution, using either a Molecular Dynamics laser scanner (with pixel sampling) or an Eikonix 78/99 CCD scanner. Raw digitized gel images are archived on high-density DAT tape (or equivalent storage media) and a greyscale video-print prepared from the raw digital image as hard-copy backup of the gel image. Gels are processed using the Kepler® software system (produced by LSB), a commercially available workstation-based software package built on

** This material (succeeding certified batches of which are available from Hoefer Scientific Instruments) has the most linear pH gradient produced by any ampholyte tested except for the Pharmacia wide range (which has an unacceptable tendency to bind high-molecular weight acidic proteins, causing them to streak).

some of the principles of the earlier TYCHO system [34-41]. Procedure PROC008 is used to yield a spotlist giving position, shape and density information for each detected spot. This procedure makes use of digital filtering, mathematical morphology techniques and digital masking to remove the background, and uses full 2-D least-squares optimization to refine the parameters of a 2-D Gaussian shape for each spot. Processing parameters and file locations are stored in a relational database, while various log files detailing operation of the automatic analysis software are archived with the reduced data. The computed resolution and level of Gaussian convergence of each gel are inspected and archived for quality control purposes.

Experiment packages are constructed using the Kepler experiment definition database to assemble groups of 2-D patterns corresponding to the experimental groups (e.g., treated and control animals). Each 2-D pattern is matched to the appropriate "master" 2-D pattern (pattern F344MST3 in the case of Fischer 344 rat liver), thereby providing linkage to the existing rodent protein 2-D databases. The software allows experiments containing hundreds of gels to be constructed and analyzed as a unit, with up to 100 gels displayed on the screen at one time for comparative purposes and multiple pages to accommodate experiments of > 1000 gels. For each treatment, proteins showing significant quantitative differences vs. appropriate controls are selected using group-wise statistical parameters (e.g., Student's *t*-test, Kepler[®] procedure STUDENT). Proteins satisfying various quantitative criteria (such as $P < 0.001$ difference from appropriate controls) are represented as highlighted spots onscreen or on computer-plotted protein maps and stored as spot populations (i.e., logical vectors) in a liver protein database. Quantitative data (spot parameters, statistical or other computed values) are stored as real-valued vectors in the database. Analysis of coregulation is performed using a Pierson product-moment correlation (Kepler procedure CORREL) to determine whether groups of proteins are coordinately regulated by any of the treatments. Such groups can be presented graphically on a protein map, and reported together with the statistical criteria used to assess the level of coregulation. Multivariate statistical analysis (e.g., principal components' analysis) is performed on data exported to SAS (SAS Institute).

2.6 Graphical data output

Graphical results are prepared in GKS and translated within Kepler[®] into output for any of a variety of devices. Linedrawing output is typically prepared as Postscript and printed on an Apple Laserwriter. Detailed maps presented here have been generated using an ultra-high-resolution Postscript-compatible Linotronic output device. Greyscale graphics are reproduced from the workstation screen using a Seikosha videoprinter. Patterns are shown in the standard orientation, with high molecular mass at the top and acidic proteins to the left.

2.7 Experiment LSBC04

In the study described here 12-week-old Charles River male F344 rats were used. Diets were prepared at LSB, based on a Purina 5755M Basal Purified Diet. Lovastatin and cholestyramine were obtained as prescription pharma-

ceuticals, ground and mixed with the diet at concentrations of 0.075% and 1%, respectively. The high cholesterol diet was Purina 5801M-A (5% cholesterol plus 1% sodium cholate in the control diet). Animal work was carried out by Microbiological Associates (Bethesda, MD). Animals were acclimatized for one week on the control diet, fed test or control diets for one week, and sacrificed on day 8. Average daily doses of lovastatin and cholestyramine in appropriate groups were 37 mg/kg/day and 5 g/kg/day, respectively, based on the weight of the food consumed. Liver samples were collected and prepared for 2-D electrophoresis according to the standard liver protocol (homogenization in 8 volumes of 9 M urea, 2% NP-40, 0.5% dithiothreitol, 2% LKB pH 9-11 carrier ampholytes, followed by centrifugation for 30 min at $80\,000 \times g$). Kidney, brain and plasma samples were frozen. Gels were run as described above, and the data was analyzed using the Kepler[®] system. Gels were scaled, to remove the effect of differences in protein loading, by setting the summed abundances of a large number of matched spots equal for each gel (linear scaling).

3 Results and discussion

3.1 The rat liver protein 2-D map

F344MST3 is a standard 2-D pattern of rat liver proteins, based on the Fischer 344 strain. This pattern was initiated from a single 2-D gel and extensively edited in an experiment comparing it to a range of protein loads, so as to include both small spots and well-resolved representations of high-abundance spots. More than 700 rat liver 2-D patterns have been matched to F344MST3 in a series of drug effects and protein characterization experiments, and numerous new spots (induced by specific drugs, for instance) have been added as a result. A modified version including additional spots present in the Sprague-Dawley outbred rat has also been developed (data not shown). Figure 1 shows a greyscale representation and Fig. 2 a schematic plot of the master pattern. More than 1200 spots are included, most of which are visible on typical gels loaded with 10 μ L of solubilized liver protein prepared by the standard method and stained with colloidal Coomassie Blue. Master spot numbers (MSN's) have been assigned to all proteins, and appear in the following figures, each showing one quadrant of the pattern. Figure 3 shows the upper left (acidic, high molecular mass) quadrant, Fig. 4 the upper right (basic, high molecular mass) quadrant, Fig. 5 the lower left (acidic, low molecular mass) quadrant, and Fig. 6 the lower right (basic, low molecular mass) quadrant. The quadrants overlap as an aid to moving between them. The gel position (in 100 micron units), isoelectric point (relative to the CPK internal *pI* standards) and SDS molecular mass (from the calibration curve in Fig. 8) are listed for each spot (Table 1). Because of the precision of the CPK-*pI* values, these parameters can be used to relate spot locations between gel systems more reliably than using *pI* measurements expressed as pH. A major objective of current studies is the identification of all major spots corresponding to known liver proteins, as well as rigorous definitions of subcellular organelle contents. Of particular interest to us is the parallel development of identifications in the rat and mouse liver maps, allowing detailed comparisons of gene expression effects in the two systems. The results of these studies will be presented systematically in a later edition of this database,

but we include here a useful series of 22 orienting identifications as an aid to other users of the rat liver pattern (Table 2).

3.2 Carbamylated charge standards, computed pI's and molecular mass standardization

We have previously shown that the use of a system of closely-spaced internal pI markers (made by carbamylating a basic protein) offers an accurate and workable solution to the problem of assigning positions in the pI dimension [32]. The same system, based on 36 protein species made by carbamylating rabbit muscle CPK, has been used here to assign pI's to most rat liver acidic and neutral proteins. The standards were coelectrophoresed with total liver proteins, and the standard spots added to a special version of the master pattern F344MST3. The gel X-coordinates of all liver protein spots lying within the CPK charge train were then transformed into CPK pI positions by interpolation between the positions of immediately adjacent standards (Table 1) using a Kepler[®] vector procedure.

It has proven possible to compute fairly accurate pI values for many proteins from the amino acid composition [42]. We have attempted here to test a further elaboration of this approach, in which we computed pI's for the CPK standards themselves, based on our knowledge of the rabbit muscle CPK sequence and the fact that adjacent members of the charge train typically differ by blockage of one additional lysine residue (Table 3). We compared these values to similar computed pI's for an additional set of carbamylated standards made from human hemoglobin beta chains and a series of rat liver and human plasma proteins of known position and sequence (Fig. 7, Table 4). The result demonstrates good concordance between these systems. Two proteins show significant deviations: liver fatty-acid binding protein (FABP; #1 in Table 4) and protein disulphide isomerase (#20 in the table). The FABP spot present on F344MST3 may represent a charge-modified version of a more basic parent spot closer to the expected pI, not resolved in the IEF/SDS gel. Of particular importance is the fact that, by comparing computed pI's of sequenced but unlocated proteins with the CPK pI's, we can assign a probable gel location without making any assumptions regarding the actual gel pH gradient. This offers a useful shortcut, given the vagaries of pH measurement on small diameter IEF gels. We have used this approach to compute the CPK pI's of all rat and mouse proteins in the PIR sequence database, as an aid to protein identification (data not shown).

In order to standardize SDS molecular weight (SDS-MW), we have used a standard curve fitted to a series of identified proteins (Fig. 8). Rather than using molecular mass *per se*, we have elected to use the number of amino acids in the polypeptide chain, as perhaps a better indication of the length of the SDS-coated rod that is sieved by the second dimension slab. The resulting values were multiplied by 112 (the weighted average mass of amino acids in sequenced proteins) to give predicted molecular masses. Because we use gradient slabs, we have not constrained the fitted curve to conform to any predetermined model; rather we tried many equations and selected the best using the program "Tablecurve" on a PC. The equation chosen was $y = a + bx + c/x^2$, where y is the number of residues, x is the gel

Y coordinate, a is 511.83, b is -0.2731 and c is 33183801. The resulting fit appears to be fairly good over a broad range of molecular mass.

3.3 An example of rat liver gene regulation: Cholesterol metabolism

Experiment LSBC04 was designed as a small-scale test of the regulation of cholesterol metabolism *in vivo* by three agents included in the diet: lovastatin (Mevacor[®], an inhibitor of HMG-CoA reductase); cholestyramine (a bile acid sequestrant that has the effect of removing cholesterol from the gut-liver recirculation); and cholesterol itself. The first two agents should lower available cholesterol and the third should raise it, allowing manipulation of relevant gene expression control systems in both directions. Such an experiment offers an interesting test of the 2-D mapping system since most of the pathway enzymes are present in low abundance, many are membrane-bound and difficult to solubilize, and the pathway itself is complex. Approximately 1000 proteins were separated and detected in liver homogenates. Twenty-one proteins were found to be affected by at least one treatment, and these could be divided into several coregulated groups.

3.3.1 MSN 413 (putative cytosolic HMG-CoA synthase) and sets of spots regulated coordinately or inversely

One group of spots (including a spot assigned to the cytosolic HMG-CoA synthase, MSN 413) showed the expected increase in abundance with lovastatin or cholestyramine, the synergistic further increase with lovastatin and cholestyramine, and a dramatic decrease with the high cholesterol diet. Spot number 413 is the most strongly regulated protein in the present experiment, showing a 5- to 10-fold induction after a 1 week treatment with 0.075% lovastatin and 1% cholestyramine in the diet (Figs. 9 and 10). Its expression follows precisely the expectation for an enzyme whose abundance is controlled by the cholesterol level; it is progressively increased from the control levels by cholestyramine, lovastatin and lovastatin plus cholestyramine, and it sinks below the threshold of detection in animals fed the high cholesterol diet. This spot has been tentatively identified as the cytosolic HMG-CoA synthase, based on a reaction with an antiserum to that protein provided by Dr. Michael Greenspan at Merck Sharp & Dohme Research Laboratories. This enzyme lies immediately before HMG-CoA reductase in the liver cholesterol biosynthesis pathway, and is known to be co-regulated with it. Spot 413 has an SDS molecular weight of about 54 000 and a CPK pI of -11.4, in reasonably close agreement with a molecular weight of 57 300 and a CPK pI of -15.7 computed from the known sequence of the hamster enzyme [43].

Using a classical product-moment correlation test (Kepler procedure CORREL), a series of five additional spots was found to be coregulated with 413. The level of correlation was exceedingly high (> 95%). Two of these, 1250 and 933, are at similar molecular weights and approximately one charge more acidic than 413 (Fig. 9), indicating that they may be covalently modified forms of the 413 polypeptide. This suspicion is strengthened by the observation that both spots are also stained by the antibody to cytosolic HMG-CoA synthase. The remaining three correlated spots appear

to comprise an additional related pair (1253 and 1001) of around 40 kDa and a single spot (1119) of around 28 kDa. Because these two presumed proteins are present at substantially lower abundances than 413, and because the cytosolic HMG-CoA synthase is reported to consist of only one type of polypeptide, they are likely to represent other, very tightly coregulated enzymes. A second group of six spots was selected based on a regulatory pattern close to the inverse of that for spot 413 (MSN's 34, 79, 178, 182, 204, 347; data not shown). For these proteins, the lowest level of expression occurs with exposure to lovastatin plus cholestyramine and the highest level upon exposure to the high-cholesterol diet. Spots 182 and 79 are highly correlated and lie about one charge apart at the same molecular weight; they may thus be isoforms of a single protein. The other four spots probably represent additional enzymes or subunits.

3.3.2 MSN 235 and coregulated spots

A third group of five spots, mainly comprised of mitochondrial proteins including putative mitochondrial HMG-CoA synthase spots, showed a modest induction by lovastatin alone, but little or no effect with any of the other treatments (including the combination of lovastatin and cholestyramine; Fig. 12). This result is intriguing because lovastatin was expected to affect only the regulation of enzymes of cholesterol synthesis, which is entirely extra-mitochondrial. Three of the spots (235, 134, 144) form a closely-packed triad at approximately 30 kDa, and are likely to represent isoforms of one protein. All three spots are stained by an antibody to the mitochondrial form of HMG-CoA synthase obtained from Dr. Greenspan. Subcellular fractionation indicates a mitochondrial location. The other two spots (633 at about 38 kDa and 724 at about 69 kDa) are each present at lower abundance than the members of the triad.

3.3.3 An example of an anti-synergistic effect

A sixth spot (367) shows strong induction by lovastatin (two- to threefold), and about half as much induction with lovastatin plus cholestyramine, but without sharing the animal-animal heterogeneity pattern of the 235-set (Fig. 13). This protein is also mitochondrial, and represents the clearest example of an anti-synergistic effect of lovastatin and cholestyramine. The existence of such an effect demonstrates that lovastatin and cholestyramine do not act exclusively through the same regulatory pathway.

3.3.4 Complexity of the cholesterol synthesis pathway

Taken together, these results suggest that treatment with lovastatin alone can affect both cytosolic and mitochondrial pathways using HMG-CoA, while cholestyramine, on the other hand, either alone or in combination with lovastatin, produces a strong effect on the putative cytosolic pathway, but little or no effect on the putative mitochondrial pathway. An explanation for this difference may lie in lovastatin's effect on levels of HMG-CoA and related precursor compounds that are exchanged between the cytosol and the mitochondrion, whereas cholestyramine should affect only the cytosolic pathways directly controlled by cholesterol and bile acid levels. It remains to be explained why some

proteins of the putative mitochondrial pathway are so much more variable in their expression in all groups. An examination of all the coregulated groups suggests that quantitative statistical techniques can extract a wealth of interesting information from large sets of reproducible gels. The abundance of spots in the 413 coregulation group, for example, shows an amazing level of concordance in their relative expression among the five individuals of the lovastatin and cholestyramine treatment group. This effect is not due to differences in total protein loading, since they have already been removed by scaling, and since proteins with quite different regulation patterns can be demonstrated (e.g., Fig. 13). Such effects raise the possibility that many gene coregulation sets may be revealed through the study of a sufficiently large population of control animals (*i.e.*, without any experimental manipulation). This approach, exploiting natural biological variation in protein expression instead of drug effects, offers an important incentive for the construction of a large library of control animal patterns.

4 Conclusions

Because of the widespread use of rat liver in both basic biochemistry and in toxicology, there is a long-term need for a comprehensive database of liver proteins. The rat liver master pattern presented here has proven to be an accurate representation of this system, having been matched to more than 700 gels to date. As the number of proteins identified and the number of compounds tested for gene expression effects grows, we expect this database to contribute valuable insights into gene regulation. Its practical utility in several areas of mechanistic toxicology is already being demonstrated.

Received September 11, 1991

5 References

- [1] O'Farrell, P., *J. Biol. Chem.* 1975, 250, 4007-4021.
- [2] Klose, J., *Humangenetik* 1975, 26, 231-243.
- [3] Scheele, G. A., *J. Biol. Chem.* 1975, 250, 5375-5385.
- [4] Iborra, G. and Buhler, J. M., *Anal. Biochem.* 1976, 74, 503-511.
- [5] Anderson, N. G. and Anderson, N. L., *Behring. Inst. Mitt.* 1979, 63, 169-210.
- [6] Anderson, N. G. and Anderson, N. L., *Clin. Chem.* 1982, 28, 739-748.
- [7] Heydorn, W. E., Creed, G. J. and Jacobowitz, D. M., *J. Pharmacol. Exp. Therap.* 1984, 229, 622-628.
- [8] Anderson, N. L., Nance, S. L., Tollaksen, S. L., Giere, F. A. and Anderson, N. G., *Electrophoresis* 1985, 6, 592-599.
- [9] Racine, R. R. and Langley, C. H., *Biochem. Genet.* 1980, 18, 185-197.
- [10] Klose, J., *Mol. Evol.* 1982, 18, 315-328.
- [11] Neel, J. V., Baier, L., Hanash, S. and Erickson, R. P., *J. Hered.* 1985, 76, 314-320.
- [12] Marshall, R. R., Raj, A. S., Grant, F. J. and Heddle, J. A., *Can. J. Genet. Cytol.* 1983, 25, 457-446.
- [13] Taylor, J., Anderson, N. L., Anderson, N. G., Gemmell, A., Giometti, C. S., Nance, S. L. and Tollaksen, S. L., in: Dunn, M. J. (Ed.), *Electrophoresis '86*, Verlag Chemie, Weinheim 1986, pp. 583-587.
- [14] Giometti, C. S., Gemmell, M. A., Nance, S. L., Tollaksen, S. L. and Taylor, J., *J. Biol. Chem.* 1987, 262, 12764-12767.
- [15] Anderson, N. L., Giere, F. A., Nance, S. L., Gemmell, M. A., Tollaksen, S. L. and Anderson, N. G., in: Galteau, M.-M. and Siest, G. (Eds.), *Progrès Récents en Electrophorèse Bidimensionnelle*, Presses Universitaires de Nancy, Nancy 1986, pp. 253-260.
- [16] Anderson, N. L., Swanson, M., Giere, F. A., Tollaksen, S., Gemmell, A., Nance, S. L. and Anderson, N. G., *Electrophoresis* 1986, 7, 44-48.

- [17] Anderson, N. L., Giere, F. A., Nance, S. L., Gemmell, M. A., Tollaksen, S. L. and Anderson, N. G., *Fundam. Appl. Toxicol.* 1987, 8, 39-50.
- [18] Anderson, N. L., in: *New Horizons in Toxicology*, Eli Lilly Symposium, 1991, in press.
- [19] Antoine, B., Rahimi-Pour, A., Siest, G., Magdalou, J. and Galteau, M. M., *Cell. Biochem. Funct.* 1987, 5, 217-231.
- [20] Elliott, B. M., Ramasamy, R., Stonard, M. D. and Spragg, S. P. *Biochim. Biophys. Acta* 1986, 870, 135-140.
- [21] Huber, B. E., Heilman, C. A., Wirth, P. J., Miller, M. J. and Thorgeirsson, S. S., *Hepatology* 1986, 6, 209-219.
- [22] Wirth, P. J. and Vesterberg, O., *Electrophoresis* 1988, 9, 47-53.
- [23] Witzmann, F. A. and Parker, D. N., *Toxicol. Lett.* 1991, 57, 29-36.
- [24] Rampersaud, A., Waxman, D. J., Ryan, D. E., Levin, W. and Walz, F. G., Jr., *Arch. Biochem. Biophys.* 1985, 243, 174-183.
- [25] Vlasuk, G. P. and Walz, F. G., Jr., *Anal. Biochem.* 1980, 105, 112-120.
- [26] Anderson, N. G. and Anderson, N. L., *Anal. Biochem.* 1978, 85, 331-340.
- [27] Anderson, N. L. and Anderson, N. G., *Anal. Biochem.* 1978, 85, 341-354.
- [28] Anderson, L., Hofmann, J.-P., Anderson, E., Walker, B. and Anderson, N. G., in: Endler, A. T. and Hanash, S. (Eds.), *Two-Dimensional Electrophoresis*, VCH Verlagsgesellschaft, Weinheim 1989, pp. 288-297.
- [29] Anderson, L., *Two-Dimensional Electrophoresis: Operation of the ISO-DALT[®] System*. Large Scale Biology Press, Washington, DC 1988, ISBN 0-945532-00-8, 170pp.
- [30] Neuhoff, V., Stamm, R. and Eibl, H., *Electrophoresis* 1985, 6, 427-448.
- [31] Neuhoff, V., Arold, N., Taube, D. and Ehrhardt, W., *Electrophoresis* 1988, 9, 255-262.
- [32] Anderson, N. L. and Hickman, B. J., *Anal. Biochem.* 1979, 93, 312-320.
- [33] Sidman, K. E., George, D. E., Barker, W. C. and Hunt, L. T., *Nucl. Acids Res.* 1988, 16, 1869-1871.
- [34] Taylor, J., Anderson, N. L., Coulter, B. P., Scandora, A. E. and Anderson, N. G., in: Radola, B. J. (Ed.), *Electrophoresis '79*, de Gruyter, Berlin 1980, pp. 329-339.
- [35] Taylor, J., Anderson, N. L. and Anderson, N. G., in: Allen, R. C. and Arnaud, P. (Eds.), *Electrophoresis '81*, de Gruyter, Berlin 1981, pp. 383-400.
- [36] Anderson, N. L., Taylor, J., Scandora, A. E., Coulter, B. P. and Anderson, N. G., *Clin. Chem.* 1981, 27, 1807-1820.
- [37] Taylor, J., Anderson, N. L., Scandora, A. E., Jr., Willard, K. E. and Anderson, N. G., *Clin. Chem.* 1982, 28, 861-866.
- [38] Taylor, J., Anderson, N. L. and Anderson, N. G., *Electrophoresis* 1983, 4, 338-345.
- [39] Anderson, N. L. and Taylor, J., in: *Proceedings of the Fourth Annual Conference and Exposition of the National Computer Graphics Association*, Chicago, June 26-30, 1983, pp. 69-76.
- [40] Anderson, N. L., Hofmann, J.-P., Gemmell, A. and Taylor, J., *Clin. Chem.* 1984, 30, 2031-2036.
- [41] Anderson, L., in: Schafer-Nielsen, C. (Ed.), *Electrophoresis '88*, VCH Verlagsgesellschaft, Weinheim 1988, pp. 313-321.
- [42] Neidhardt, F. C., Appleby, D. A., Sankar, P., Hutton, M. E. and Phillips, T. A., *Electrophoresis* 1989, 10, 116-121.
- [43] Gil, G., Goldstein, J. L., Slaughter, C. A. and Brown, M. S., *J. Biol. Chem.* 1986, 261, 3710-3716.

6 Addendum 1: Figures 1-13

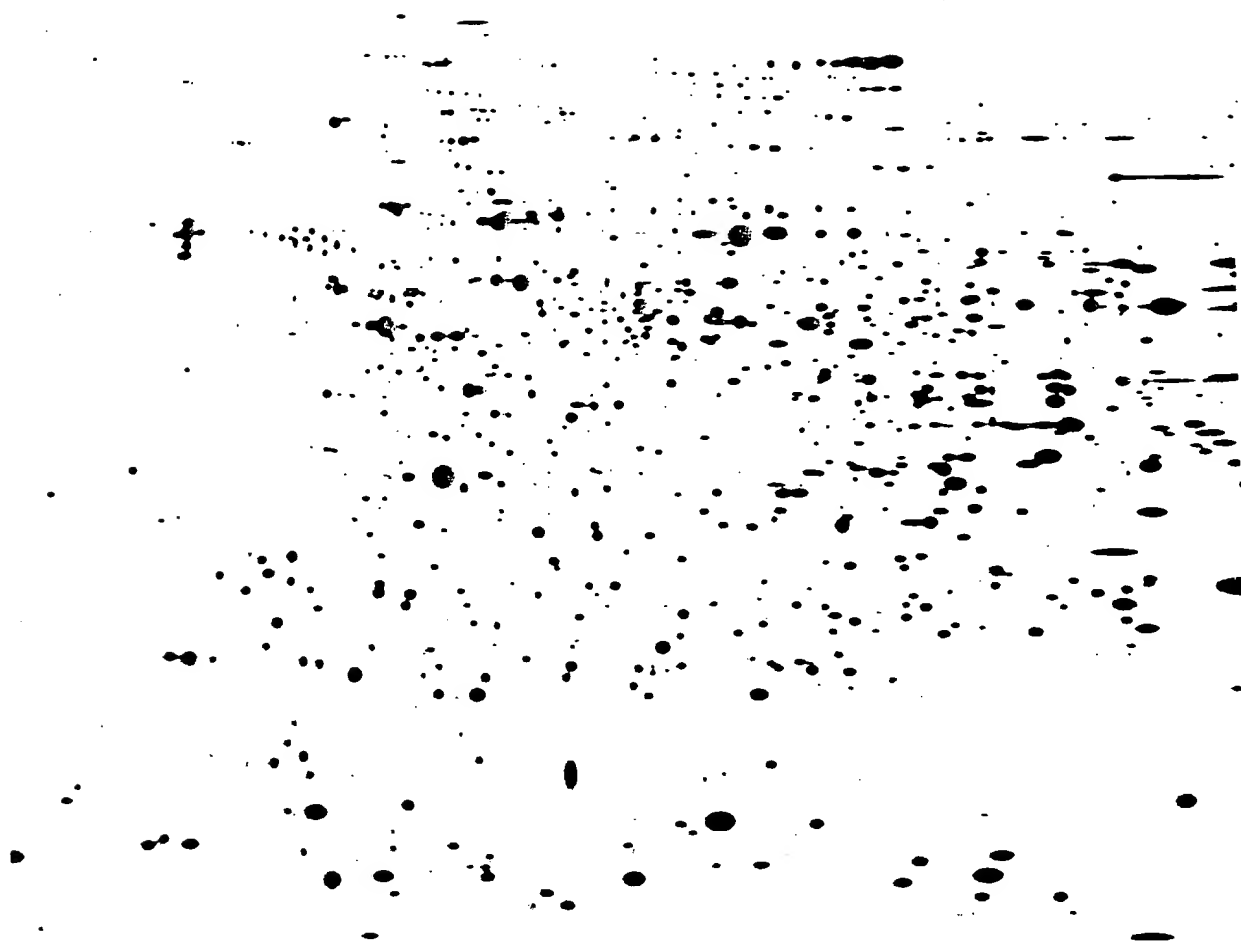


Figure 1. Synthetic representation of the standard rat liver 2-D master pattern, rendered as a greyscale image using a videoprinter.

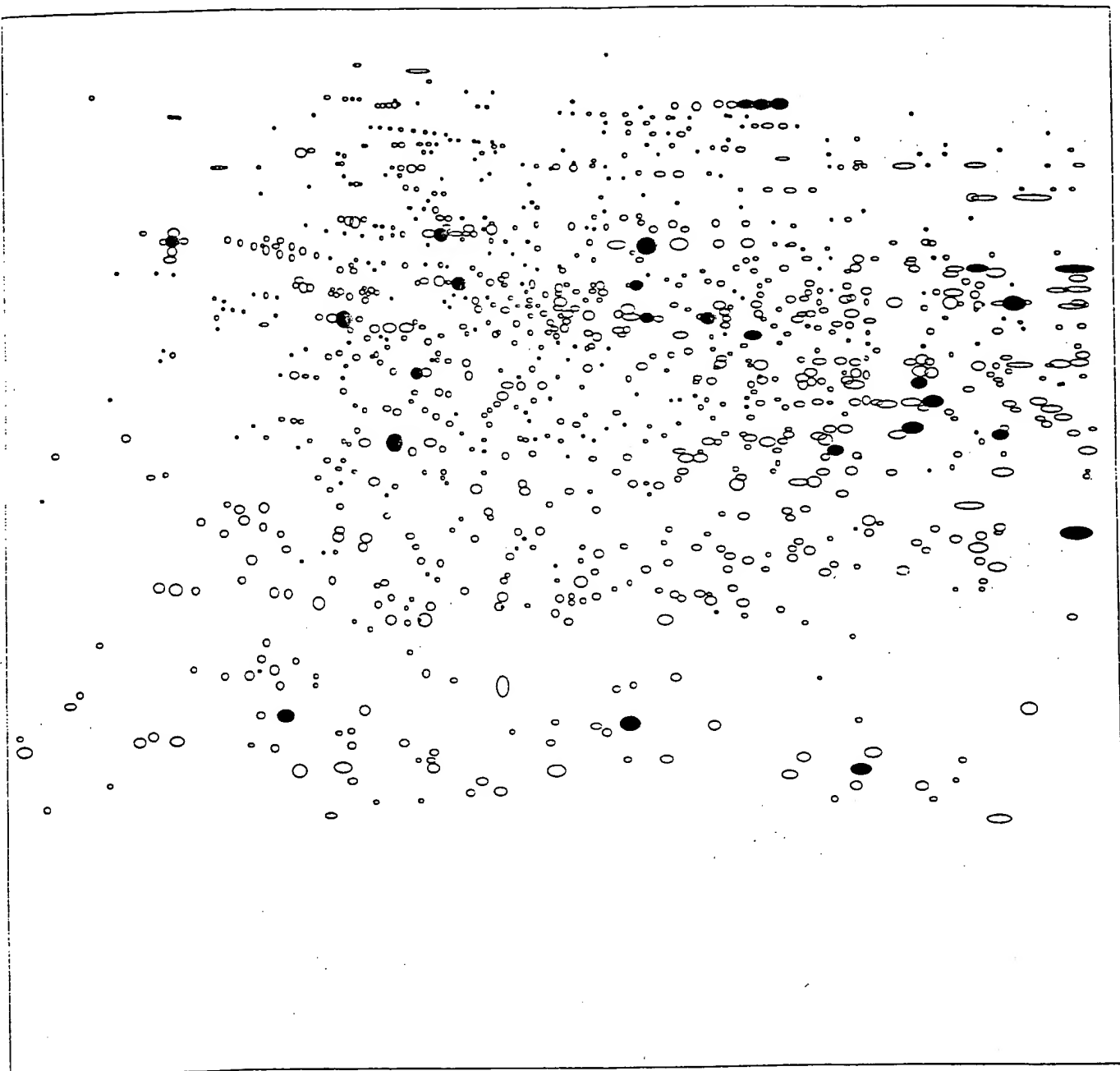


Figure 2. Schematic representation of the master pattern (the same as Fig. 1), useful as an aid in relating specific areas of Fig. 1 and the following detailed quadrants.

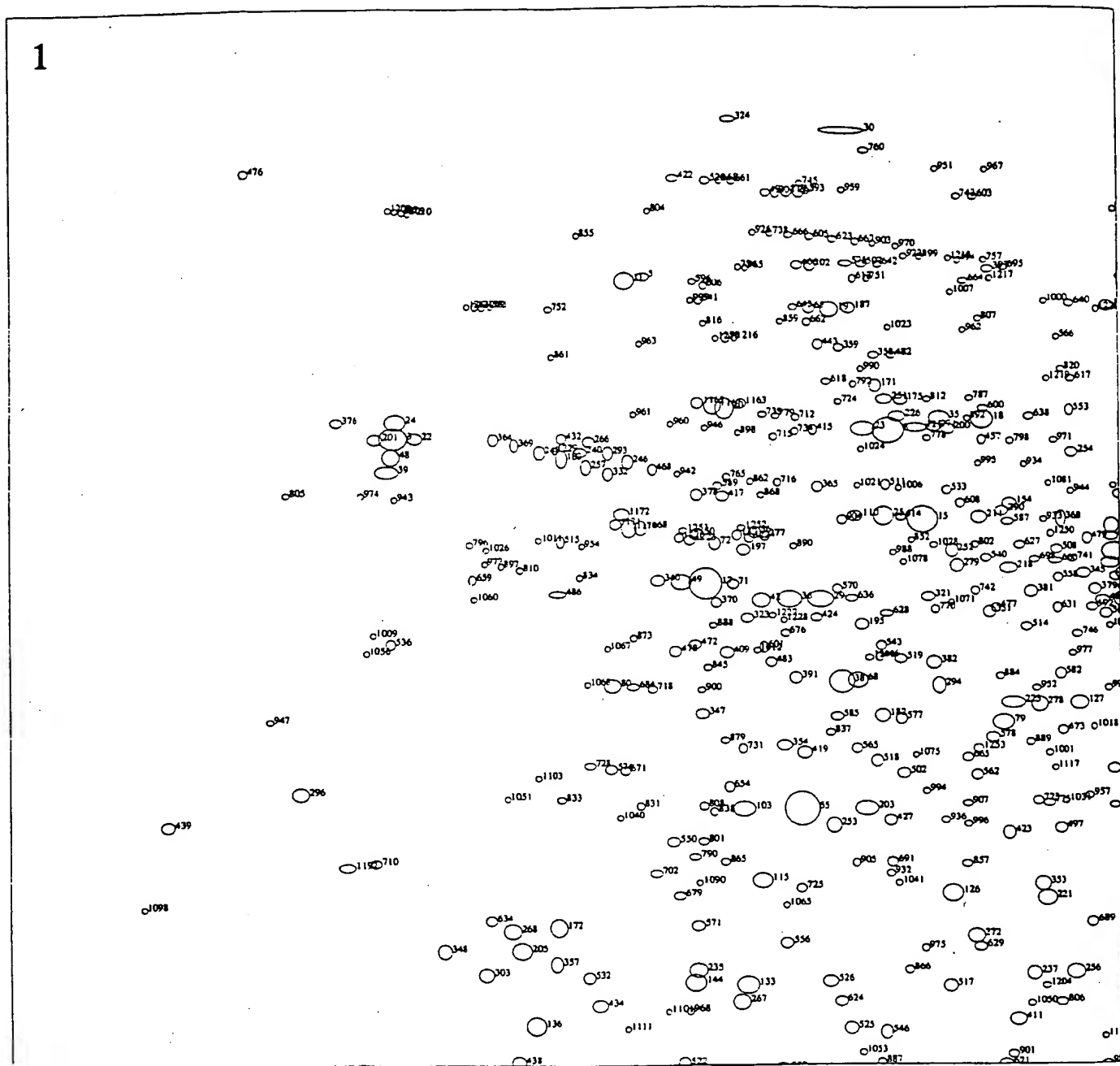


Figure 3. Upper left (high molecular weight, acidic) quadrant (#1) of the rat liver map, showing spot numbers.

2

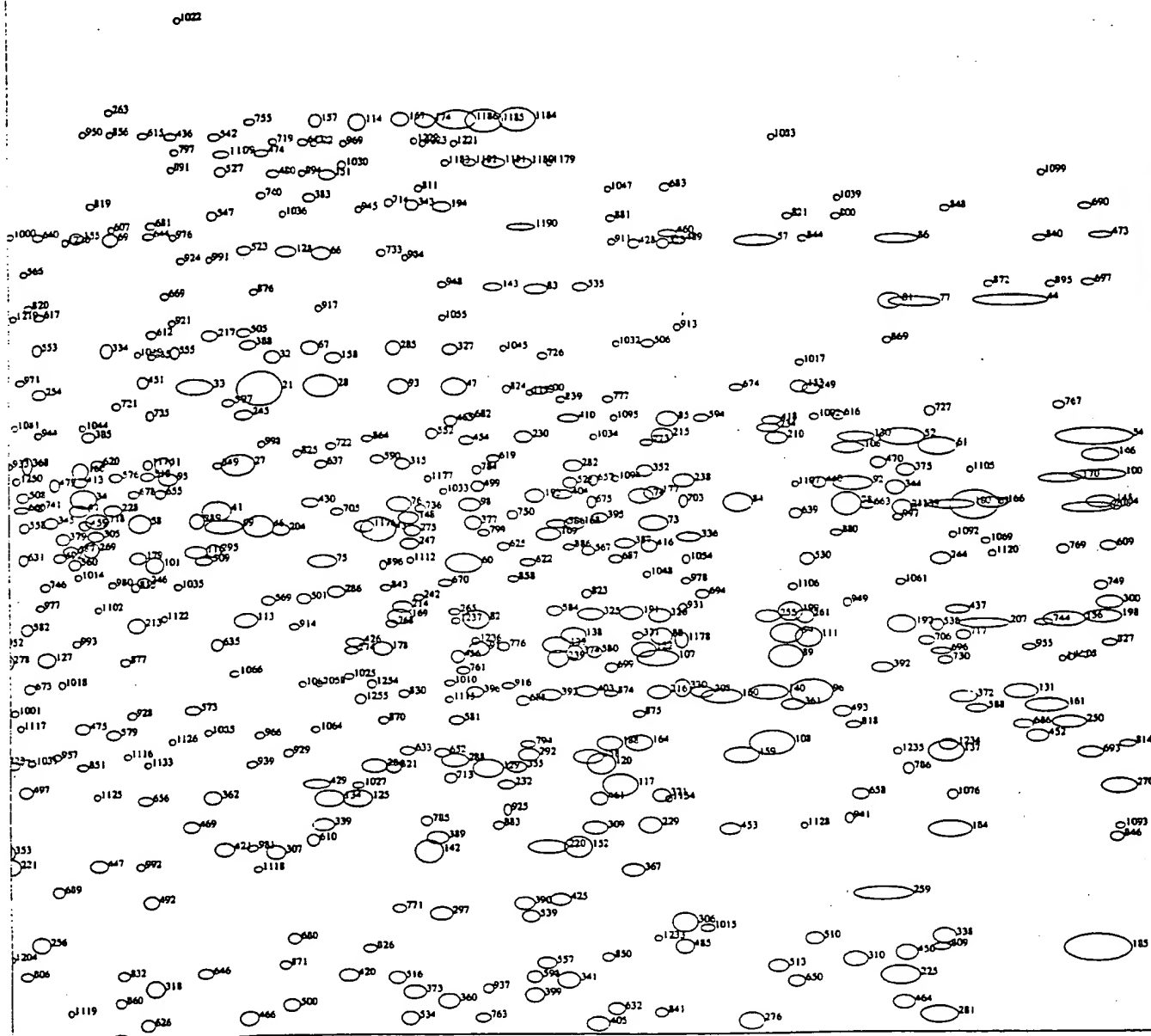


Figure 4. Upper right (high molecular weight, basic) quadrant (#2) of the rat liver map, showing spot numbers.

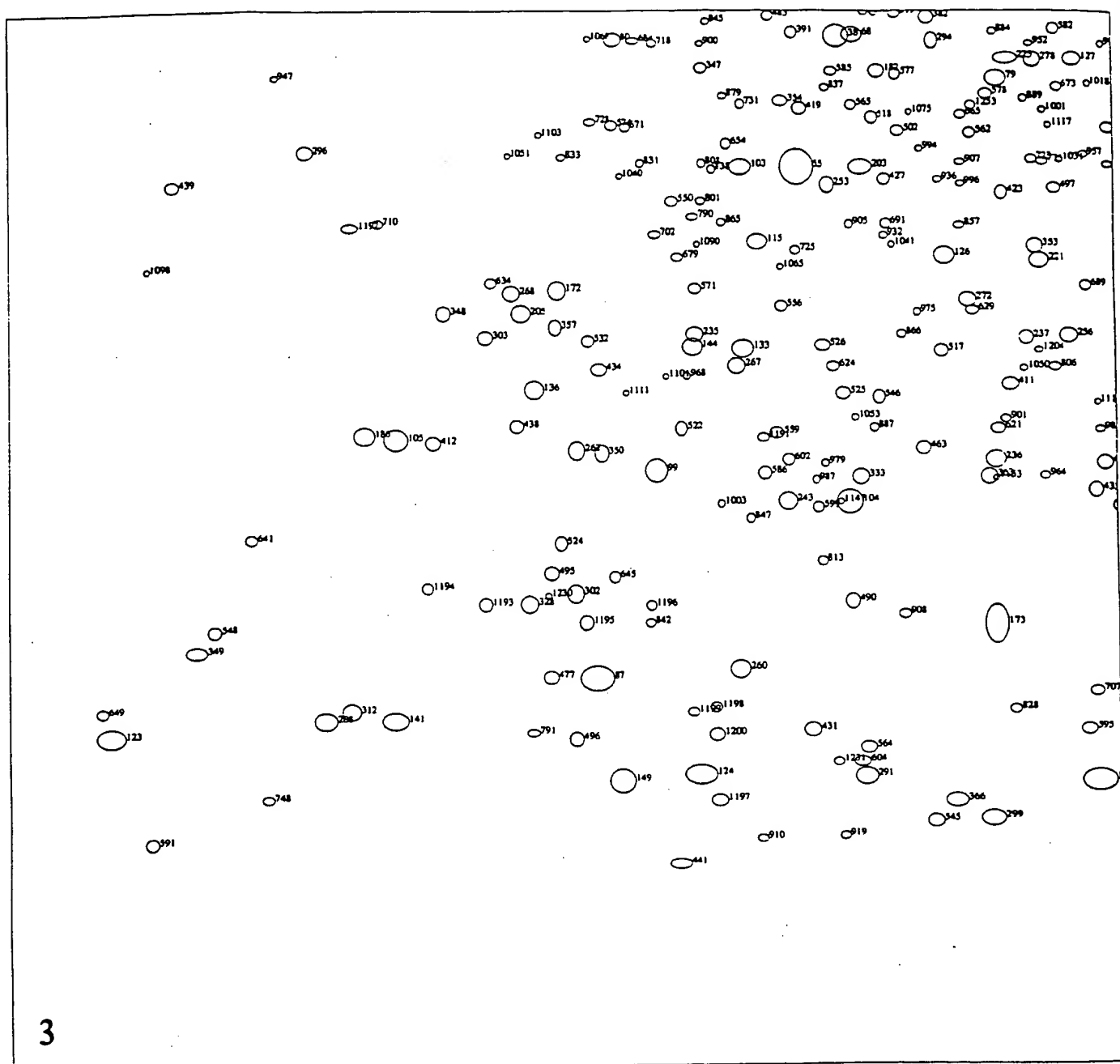


Figure 5. Lower left (low molecular weight, acidic) quadrant (#3) of the rat liver map, showing spot numbers.

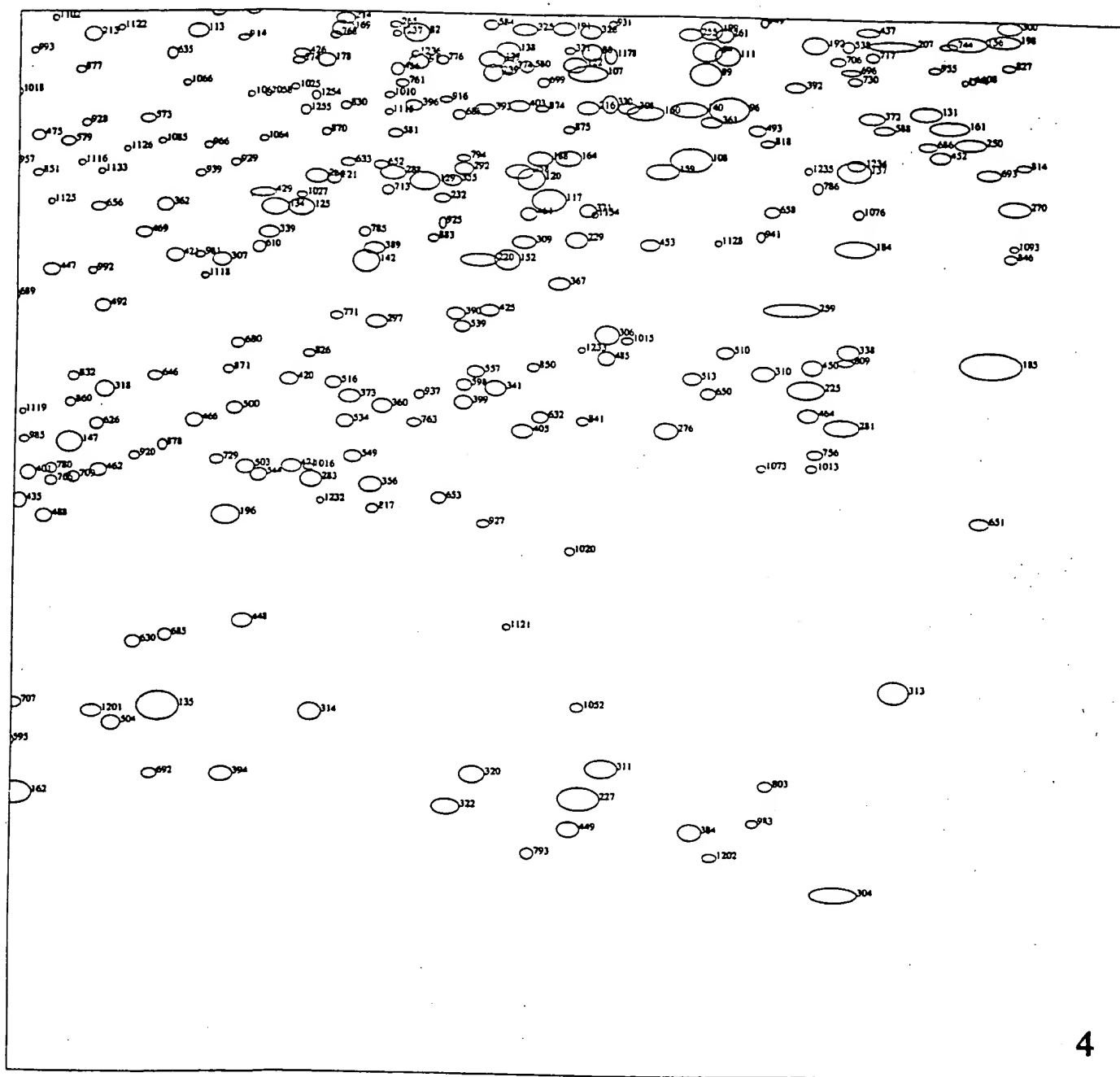


Figure 6. Lower right (low molecular weight, basic) quadrant (#4) of the rat liver map, showing spot numbers.

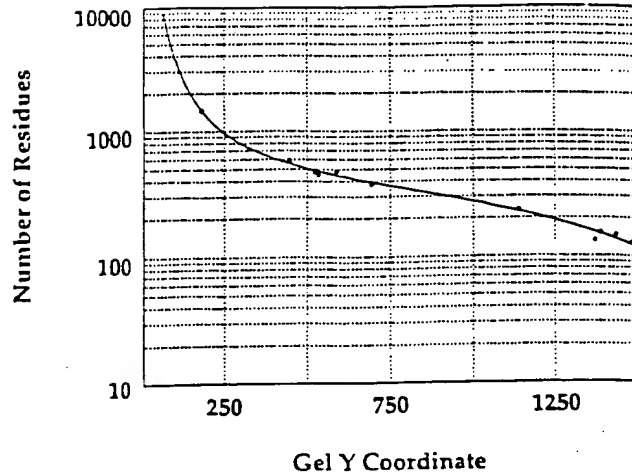
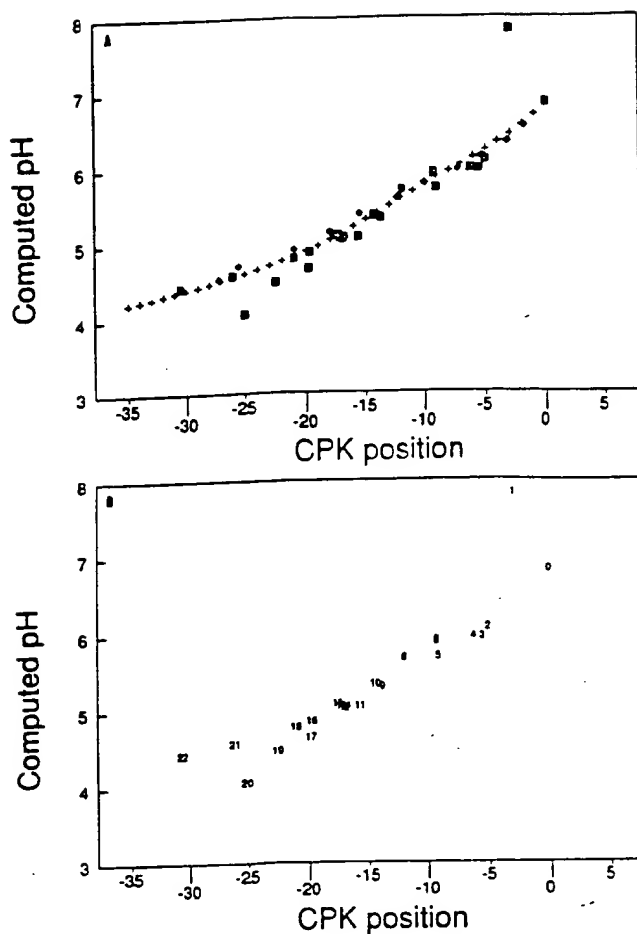


Figure 8. Plot of number of amino acids versus gel Y-position, with fitted curve used to predict molecular mass of unidentified proteins.

Figure 7. (a) Plot of computed isoelectric point versus gel X-position for two sets of carbamylated standard proteins (rabbit muscle CPK [+]) and human hemoglobin β chain, filled diamonds) and several other proteins (shaded squares). (b) The identities of the various proteins represented by the squares are indicated by the numbers in corresponding positions on (a); these refer to Table 4.

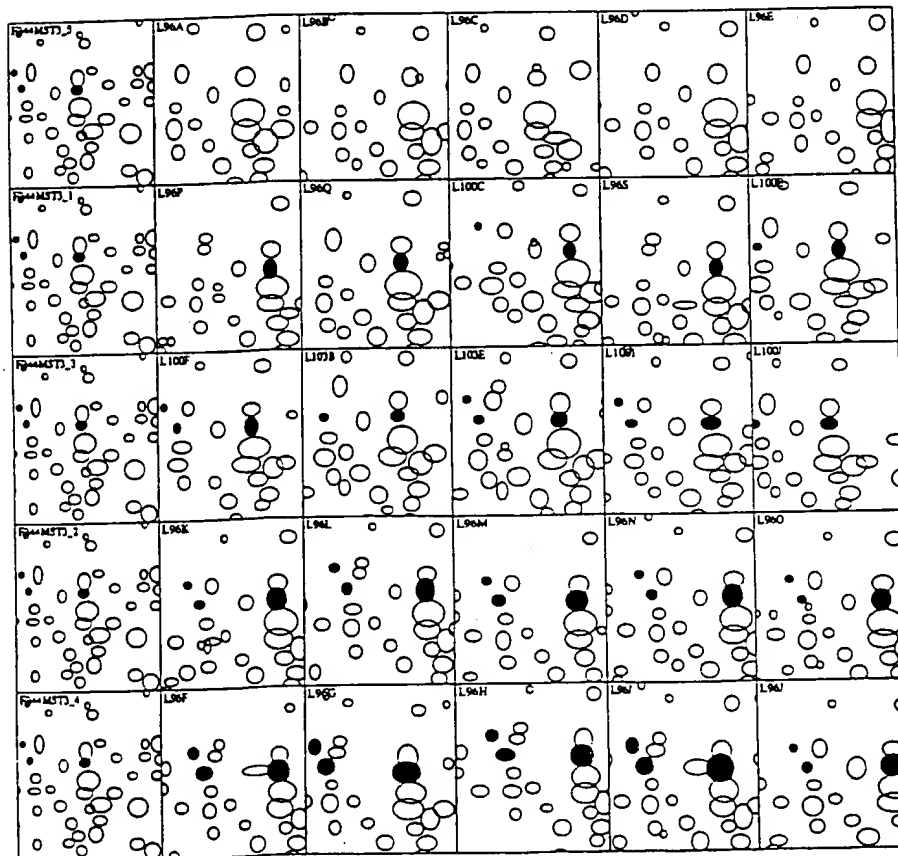


Figure 9. Montage showing effects in the region of MSN:413. The montage shows a small window into one portion of the 2-D pattern, one row of windows for each experimental group, and one panel for each gel in the experiment. The left-most pattern in each row is a group-specific copy of the master pattern followed by the patterns for the five individual rats in the group. The highlighted protein spots (filled circles) are spot 413 (on the right of each panel), identified as cytosolic HMG-CoA synthase and two modified forms of it (1250 and 933). From the top, the rows (experimental groups) are: high cholesterol, controls, cholestyramine, lovastatin, and lovastatin plus cholestyramine.

Regulation of Rat Liver 413

(Putative Cytosolic HMG-CoA Synthase, 53kd)
Test Compounds in Diet

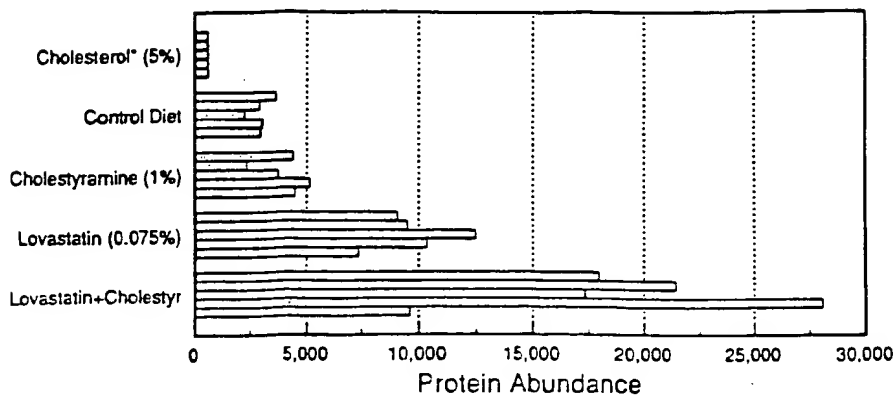


Figure 10. Bargraph showing the quantitative effects of various treatments on the abundance of MSN:413 (cytosolic HMG-CoA synthase) in the gels of Fig. 9.

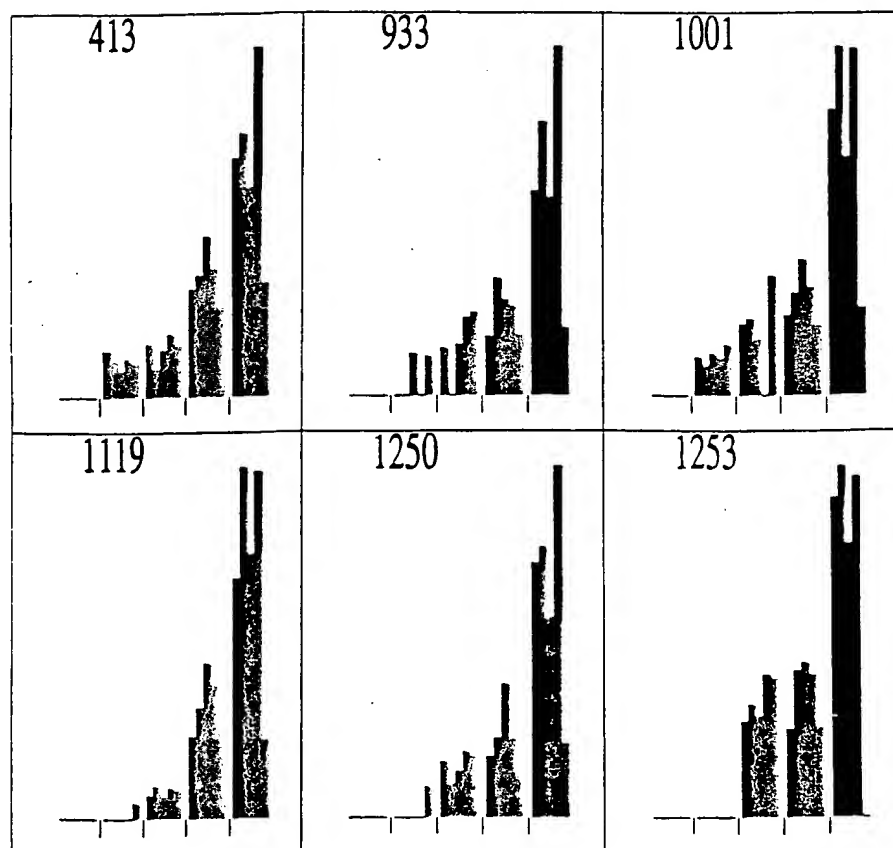


Figure 11. Bargraphs of a series of six coregulated spots including MSN:413. In the bargraphs, the abundances of the appropriate spot (master spot number shown at the top of the panel) in each animal are shown. The five five-animal groups are in the order (left to right): high cholesterol, controls, cholestyramine, lovastatin, and lovastatin plus cholestyramine. Each bar within a group represents one experimental animal liver (one 2-D gel). Note the correlated expression of the 6 spots, especially in the two far right (most strongly induced) groups.

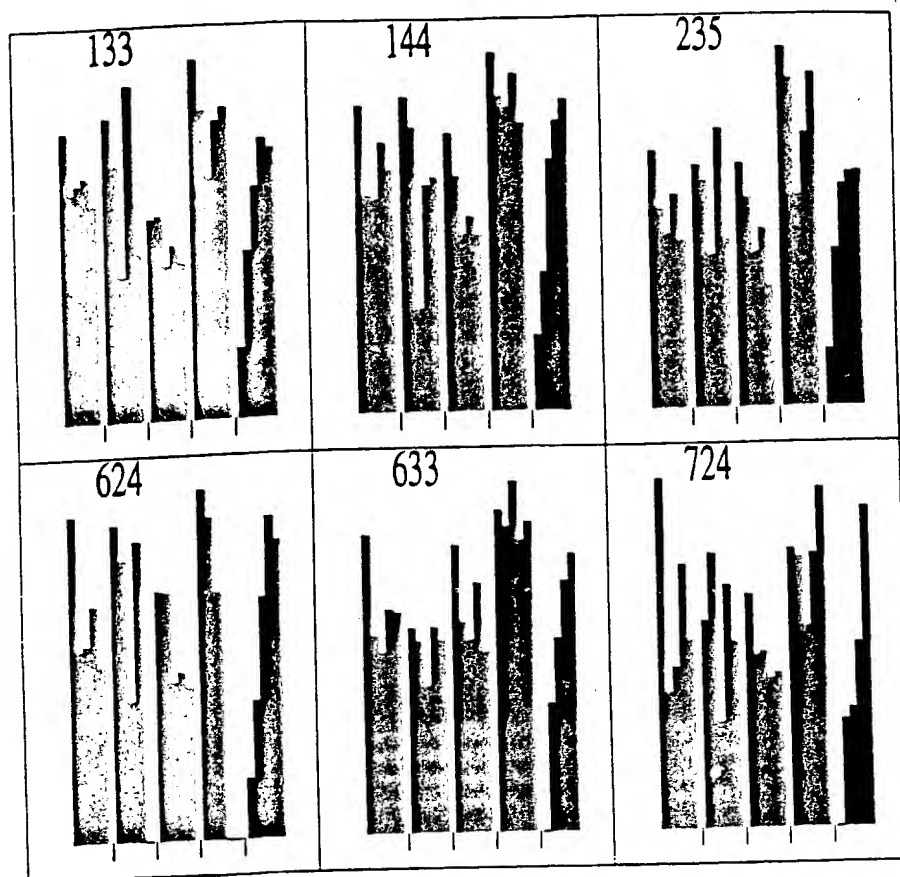


Figure 12. Data on a second coregulated group of spots, presented as in Fig. 11. The fourth experimental group (lovastatin) shows a modest induction, while the fifth group (lovastatin plus cholestyramine) does not.

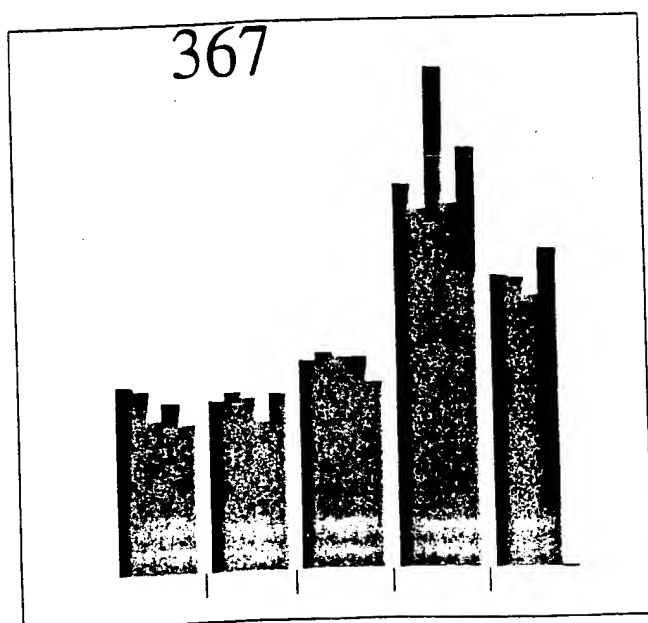


Figure 13. Data on spot MSN:367, presented as in Fig. 11. This protein shows unambiguously the anti-synergistic effect of lovastatin and cholestyramine (fifth group) as compared to lovastatin (fourth group). This response contrasts strongly with the regulation pattern seen in Fig. 11.

Table 1. Master table of proteins in the rat liver database^{a)}

MSN	X	Y	CPKpI	SDSMW	MSN	X	Y	CPKpI	SDSMW	MSN	X	Y	CPKpI	SDSMW
3	311	434	<-35.0	63,800	95	1119	536	-9.9	53,800	174	1364	183	-6.7	162,900
5	568	263	-24.3	102,900	96	1731	756	-2.0	40,700	175	825	383	-15.7	69,300
8	812	426	-16.0	64,800	97	1033	566	-11.4	51,600	177	1582	553	-3.6	52,600
11	549	268	-25.2	101,000	98	1406	565	-6.1	51,700	178	1321	710	-7.2	43,000
15	845	520	-15.3	55,200	99	578	1149	-23.8	25,000	179	1089	615	-10.4	48,300
17	629	589	-21.6	50,000	100	2004	538	>0.0	53,700	180	1866	567	-0.5	51,600
18	906	414	-14.0	66,300	101	1106	623	-10.1	47,900	181	411	295	-32.1	91,200
19	755	298	-17.5	90,200	102	482	455	-28.5	61,300	182	804	730	-16.2	42,000
20	649	403	-20.9	67,900	103	665	830	-20.2	37,300	184	1860	896	-0.6	34,500
21	1204	448	-8.7	62,100	104	773	1182	-17.0	23,800	185	1997	1017	>0.0	29,800
22	332	434	<-35.0	63,800	105	312	1117	<-35.0	26,100	186	279	1113	<-35.0	26,300
23	787	424	-16.6	65,000	106	1769	509	-1.5	56,100	187	773	296	-17.0	90,800
24	313	417	<-35.0	66,000	107	1585	720	-3.6	42,500	188	1538	807	-4.2	38,400
25	807	516	-16.1	55,500	108	1692	807	-2.4	38,300	191	1560	674	-3.9	44,900
27	1184	524	-9.0	54,900	109	1482	593	-4.8	49,700	192	1818	687	-0.9	44,200
28	1263	446	-8.0	62,400	110	778	516	-16.9	55,500	193	1469	555	-5.0	52,400
29	743	605	-17.8	49,000	111	1728	700	-2.0	43,500	194	1380	266	-6.4	101,600
30	768	112	-17.2	348,600	113	1191	680	-8.9	44,500	195	784	632	-16.7	47,300
32	1216	417	-8.6	66,000	114	1298	185	-7.5	160,800	196	1227	1185	-8.4	23,700
33	1145	445	-9.5	62,500	115	682	907	-19.6	34,100	197	667	553	-20.1	52,600
34	1037	555	-11.3	52,400	116	1146	810	-8.5	48,700	198	2006	681	>0.0	44,500
35	863	412	-14.9	66,600	117	1548	849	-4.1	36,500	199	1711	674	-2.2	44,900
36	712	606	-18.7	48,900	118	1050	577	-11.1	50,800	200	872	424	-14.7	65,000
38	763	694	-17.3	43,800	120	1530	828	-4.3	37,400	201	292	435	<-35.0	63,700
39	304	470	<-35.0	59,800	121	638	423	-15.4	65,200	202	736	253	-18.0	107,800
41	1165	569	-9.2	51,400	122	1572	712	-3.8	42,900	203	786	829	-16.7	37,400
42	684	607	-19.6	46,800	123	23	1433	<-35.0	15,300	204	1224	589	-8.5	50,000
43	1318	589	-7.3	50,000	124	621	1474	-21.9	13,900	205	439	983	-30.9	31,100
44	1924	362	-0.1	74,600	125	1298	862	-7.5	36,000	206	1994	571	>0.0	51,300
46	1203	586	-8.7	50,200	126	872	921	-14.7	33,500	207	1895	687	-0.3	44,200
47	1391	447	-6.3	62,300	127	1000	717	-12.0	42,600	208	240	1418	<-35.0	15,800
48	309	454	<-35.0	61,500	128	1229	311	-8.4	86,100	210	1700	499	-2.3	57,000
49	605	587	-22.5	50,100	129	1422	832	-5.8	37,300	211	902	517	-14.1	55,400
50	621	535	-21.8	53,900	130	1776	499	-1.4	57,000	213	1087	684	-10.4	44,400
51	1113	522	-10.0	55,000	131	1930	757	-0.1	40,700	214	1340	668	-7.0	45,200
52	1820	499	-0.9	57,000	132	660	537	-20.4	53,800	215	1591	495	-3.5	57,300
53	725	177	-18.3	170,800	133	666	1019	-20.2	29,700	216	1585	755	-3.6	40,700
54	2001	500	>0.0	56,900	134	1271	862	-7.9	36,000	217	1159	393	-9.3	69,300
55	722	830	-18.4	37,300	135	1161	1389	-9.3	16,800	218	931	572	-13.5	51,200
56	678	533	-19.8	54,100	136	453	1063	-29.7	28,100	219	713	177	-18.7	170,500
57	1682	302	-2.5	89,000	137	1858	823	-0.6	37,700	220	1479	911	-4.9	33,900
58	1091	580	-10.3	50,600	138	1504	697	-4.6	43,700	221	965	927	-12.8	33,300
59	1171	585	-9.2	50,300	139	1488	707	-4.8	43,200	223	934	716	-13.5	42,700
60	1400	624	-6.2	47,800	140	1689	756	-2.4	40,700	225	1812	1045	-1.0	28,800
61	1853	508	-0.6	56,200	141	311	1417	<-35.0	15,800	226	821	411	-15.8	66,800
62	1888	567	-0.4	51,500	142	1366	915	-6.7	33,800	227	1586	1483	-3.6	13,600
65	735	297	-18.1	90,500	143	1429	346	-5.7	77,900	228	1065	567	-10.8	51,600
66	1263	312	-8.0	85,900	144	615	1017	-22.1	29,800	229	1577	890	-3.7	34,800
67	1252	407	-8.1	67,300	145	2006	566	>0.0	51,600	230	1458	496	-5.2	57,300
68	779	682	-16.8	43,900	146	2006	518	>0.0	55,300	232	1440	849	-5.5	36,500
69	1064	296	-10.8	90,800	147	1070	1108	-10.7	26,500	234	1692	489	-2.4	57,900
71	656	589	-20.6	50,000	148	1347	578	-6.9	50,800	235	618	1004	-22.0	30,300
72	638	545	-21.2	53,100	149	541	1481	-25.7	13,700	236	920	1138	-13.7	25,400
73	1582	583	-3.6	50,400	150	1645	760	-2.8	40,500	237	952	1008	-13.1	30,200
74	1570	556	-3.8	52,300	151	1269	236	-7.9	117,000	238	1611	541	-3.2	53,500
75	1264	621	-8.0	48,000	152	1507	911	-4.5	33,900	239	1489	720	-4.8	42,500
76	1338	564	-7.0	51,800	153	1722	448	-2.1	62,100	240	501	448	-27.7	62,100
77	1833	363	-0.8	74,400	154	932	503	-13.5	56,600	241	1820	569	-0.9	51,400
78	1767	565	-1.5	51,700	155	1031	294	-11.4	91,400	242	1357	658	-6.8	45,800
79	925	738	-13.6	41,600	156	1970	684	>0.0	44,400	243	711	1182	-18.7	23,800
80	534	698	-26.1	43,600	157	1258	183	-8.1	162,400	244	1855	621	-0.6	48,000
81	1811	363	-1.0	74,500	158	1275	417	-7.8	65,900	245	1189	474	-8.9	59,300
82	1412	681	-6.0	44,500	159	1663	820	-2.6	37,800	246	551	459	-25.1	61,000
83	1471	347	-5.0	77,500	160	1034	527	-11.4	54,600	247	1348	604	-6.9	49,100
84	1662	563	-2.7	51,800	161	1953	771	>0.0	40,000	248	460	448	-29.3	62,100
85	1596	479	-3.4	58,900	162	1020	1482	-11.6	13,700	249	1733	451	-1.9	61,800
86	1817	301	-0.9	89,100	164	1566	806	-3.8	38,400	250	1974	788	>0.0	39,200
87	516	1371	-27.0	17,400	166	1905	565	-0.2	51,700	251	808	392	-16.1	69,500
88	1589	698	-3.5	43,600	167	1340	181	-7.0	164,900	252	874	553	-14.6	52,500
89	1706	719	-2.2	42,500	168	1506	583	-4.6	50,400	253	753	848	-17.6	36,500
90	651	329	-20.8	81,700	169	1338	678	-7.0	44,700	254	995	450	-12.1	61,900
91	1415	710	-6.0	43,000	170	1969	541	>0.0	53,500	255	1690	679	-2.4	44,600
92	1773	545	-1.4	53,200	171	800	378	-16.3	71,800	256	994	1006	-12.1	30,200
93	1338	446	-7.0	62,300	172	476	958	-28.7	32,100	257	508	464	-27.4	60,400
94	1708	696	-2.2	43,700	173	919	1314	-13.7	19,300	258	1517	820	-4.4	37,800

^{a)} Master table of proteins in the rat liver database, showing spot master number, gel position (x and y), isoelectric point relative to CPK standards, and predicted molecular mass (from the standard curve of Fig. 8).

MSN	X	Y	CPKdI	SDSMW	MSN	X	Y	CPKdI	SDSMW	MSN	X	Y	CPKdI	SDSMW
259	1796	961	-1.1	31,900	345	1006	578	-11.9	50,800	426	1296	704	-7.8	43,300
260	661	1361	-20.4	17,700	346	1095	640	-10.3	46,800	427	810	843	-16.0	36,800
261	1725	679	-2.0	44,600	347	625	728	-21.7	42,000	428	1565	303	-3.9	86,700
262	496	1127	-28.0	25,800	348	361	963	-35.3	31,100	429	1259	847	-8.0	36,800
263	1063	172	-10.9	177,400	349	110	1343	<-35.0	18,300	430	1253	562	-8.1	51,900
265	1390	673	-6.3	45,000	350	521	1130	-26.7	25,700	431	734	1426	-18.1	15,500
266	510	437	-27.3	63,400	351	912	619	-13.9	48,100	432	483	433	-28.5	63,900
267	660	1038	-20.4	29,000	352	1574	530	-3.7	54,300	434	518	1041	-26.9	26,900
268	430	961	-31.0	31,900	353	961	912	-12.9	33,900	435	1020	1170	-11.6	24,300
269	1044	606	-11.2	48,900	354	706	762	-18.9	40,400	436	1122	196	-9.8	147,600
270	2019	853	>0.0	36,300	355	1450	830	-5.3	37,300	437	1870	673	-0.5	45,000
271	857	422	-15.0	65,200	356	1374	1152	-6.5	24,900	438	435	1102	-31.0	26,700
272	895	968	-14.2	31,700	357	474	997	-28.7	30,600	439	86	847	<-35.0	36,600
274	1292	712	-7.6	42,900	358	798	346	-16.3	77,800	440	1740	544	-1.8	53,200
275	1350	590	-6.9	49,900	359	764	338	-17.3	79,400	441	599	1571	-22.8	10,800
276	1670	1089	-2.6	27,100	360	1364	1068	-6.4	27,900	443	743	335	-17.8	80,100
277	688	538	-19.4	53,700	361	1713	769	-2.1	40,100	446	801	668	-16.2	45,200
278	961	718	-13.0	42,600	362	1161	859	-9.3	36,100	447	1050	926	-11.1	33,300
279	879	570	-14.5	51,300	363	914	1156	-13.8	24,800	448	1245	1298	-8.2	19,800
281	1648	1084	-0.7	27,300	364	412	435	-32.0	63,700	449	1576	1516	-3.7	12,600
282	1505	525	-4.6	54,800	365	741	486	-17.9	58,200	450	1818	1021	-0.9	29,600
283	1313	1147	-7.3	25,100	366	878	1503	-14.6	13,000	451	1094	440	-10.3	63,100
284	1314	829	-7.3	37,400	367	1560	935	-3.9	33,000	452	1945	802	>0.0	38,600
285	1332	408	-7.1	67,200	368	983	520	-12.4	55,200	453	1652	894	-2.8	34,600
286	1277	652	-7.8	46,100	369	434	441	-31.0	63,000	454	1403	500	-6.1	56,900
288	1391	824	-6.3	37,600	370	639	610	-21.2	48,700	456	1394	718	-6.3	42,600
289	1147	579	-9.5	50,700	371	1587	860	-3.6	36,100	457	905	436	-14.0	63,500
290	925	511	-13.6	55,900	372	1875	762	-0.5	40,400	459	1038	581	-11.3	50,500
291	787	1476	-16.6	13,900	373	1351	1059	-6.8	28,300	460	1598	294	-3.4	91,400
292	1462	818	-5.1	37,800	374	1506	715	-4.6	42,700	461	1528	863	-4.3	35,900
293	531	449	-26.3	62,000	375	1823	532	-0.9	54,200	462	1098	1137	-10.2	25,400
294	860	698	-14.9	43,600	376	254	417	<-35.0	65,900	463	849	1125	-15.2	25,800
295	1162	609	-9.3	48,700	377	1409	583	-6.1	50,400	464	1814	1072	-0.9	27,800
296	218	814	<-35.0	38,000	378	621	494	-21.8	57,500	465	1388	481	-6.3	58,700
297	1377	979	-6.5	31,300	379	1017	595	-11.7	49,600	466	1194	1084	-8.9	27,300
299	913	1523	-13.9	12,400	381	953	598	-13.1	49,400	468	577	467	-23.9	60,100
300	2012	667	>0.0	45,300	382	856	674	-15.0	44,900	469	1140	888	-9.6	34,900
301	702	178	-19.0	169,200	383	1252	258	-8.1	105,300	470	1797	524	-1.1	54,800
302	494	1280	-28.1	20,400	384	1699	1518	-2.3	12,500	471	1293	1133	-7.6	25,500
303	403	1008	-32.6	30,100	385	1042	493	-11.2	57,500	472	618	655	-21.9	46,000
304	1843	1585	-0.7	10,300	386	1490	583	-4.7	50,400	473	2009	299	>0.0	89,900
305	1049	593	-11.1	49,800	387	1554	603	-4.0	49,100	474	1205	215	-8.7	131,300
306	1608	989	-3.3	30,900	388	1193	404	-8.9	67,700	475	1035	788	-11.4	39,200
307	1219	916	-8.5	33,700	389	1374	902	-6.5	34,300	476	160	155	<-35.0	207,600
308	1627	755	-3.0	40,700	390	1456	969	-5.2	31,700	477	469	1370	-28.9	17,400
309	1524	892	-4.4	34,700	391	718	690	-18.5	44,000	478	599	662	-22.8	45,600
310	1769	1028	-1.5	29,400	392	1799	732	-1.1	41,900	479	1009	540	-11.8	53,500
311	1609	1451	-3.3	14,700	393	1482	758	-4.8	40,600	480	1216	235	-8.6	117,400
312	266	1408	<-35.0	16,100	394	1227	1461	-8.4	14,400	482	816	346	-15.9	77,800
313	1902	1365	-0.3	17,600	395	1530	577	-4.3	50,800	483	693	673	-19.3	44,900
314	1316	1395	-7.3	16,600	396	1410	755	-6.0	40,800	485	1608	1013	-3.3	30,000
315	1341	523	-7.0	54,900	397	912	256	-13.9	106,400	486	478	599	-28.6	49,300
318	1104	1053	-10.1	28,500	399	1465	1063	-5.0	28,100	487	1025	607	-11.5	48,800
320	1480	1459	-4.9	14,400	400	1473	450	-4.9	61,900	488	1045	1186	-11.2	23,700
321	850	603	-15.1	49,100	401	1029	1140	-11.5	25,300	489	1609	301	-3.3	89,200
322	1454	1494	-5.3	13,300	403	1516	754	-4.4	40,800	490	775	1289	-17.0	20,100
323	670	626	-20.0	47,700	404	1495	554	-4.7	52,500	491	692	178	-19.3	169,300
324	655	101	-20.6	420,500	405	1525	1092	-4.3	27,100	492	1100	964	-10.2	31,800
325	1521	675	-4.4	44,800	406	723	252	-18.4	108,000	493	1760	776	-1.6	39,700
326	1587	677	-3.6	44,700	409	650	663	-20.8	45,500	494	882	247	-14.5	110,700
327	1388	409	-6.3	67,000	410	1501	478	-4.6	59,000	495	470	1258	-28.9	21,200
328	448	1291	-30.0	20,100	411	936	1057	-13.4	28,300	496	494	1436	-28.1	15,200
330	1608	751	-3.3	40,900	412	350	1120	-35.9	26,000	497	980	852	-12.5	36,400
331	1566	697	-3.8	43,700	413	1033	538	-11.4	53,700	499	1414	546	-6.0	53,100
332	531	471	-26.3	59,600	415	737	425	-18.0	64,900	500	1234	1072	-8.3	27,800
333	784	1156	-16.7	24,700	416	1578	606	-3.7	48,900	501	1246	659	-8.2	45,700
334	1059	407	-10.9	67,300	417	646	496	-21.0	57,300	502	824	792	-15.7	39,000
335	1593	303	-3.5	86,500	418	1695	482	-2.3	58,600	503	1246	1134	-8.2	25,500
336	1616	598	-3.2	49,400	419	725	770	-18.3	40,000	504	1115	1407	-9.9	16,200
338	1854	1004	-0.6	30,300	420	1289	1041	-7.7	28,900	505	1189	391	-8.9	69,700
339	1265	888	-8.0	34,900	421	1171	912	-9.1	33,900	506	1578	402	-3.7	68,000
340	581	585	-23.6	50,300	422	599	162	-22.8	193,700	507	787	250	-16.6	109,000
341	1497	1047	-4.7	26,700	423	929	856	-13.6	36,200	508	979	552	-12.5	52,600
343	1351	265	-6.8	102,200	424	739	625	-17.9	47,700	509	1153	619	-9.4	48,100
344	1813	549	-0.9	52,800	425	1490	965	-4.7	31,800	510	1730	1006	-2.0	30,200

MSN	X	Y	CPKdI	SDSMW	MSN	X	Y	CPKdI	SDSMW	MSN	X	Y	CPKdI	SDSMW
511	809	484	-16.0	58,400	586	619	269	-21.9	100,500	674	1661	448	-2.7	62,100
512	1099	533	-10.2	54,100	567	1176	461	-9.1	60,700	675	1523	562	-4.4	51,900
513	1696	1034	-2.3	29,200	598	1465	1044	-5.0	28,800	676	708	642	-18.8	46,700
514	948	636	-13.2	47,100	599	741	1188	-17.9	23,600	677	919	615	-13.7	48,300
515	481	543	-28.5	53,400	600	907	402	-14.0	68,000	678	1085	551	-10.5	52,700
516	1334	1044	-7.1	26,800	601	687	658	-19.5	45,800	679	600	923	-22.7	33,400
517	868	1021	-14.8	29,700	602	712	1138	-18.7	25,400	680	1237	1004	-8.3	30,300
518	798	779	-16.3	39,600	603	898	181	-14.1	165,200	681	1103	283	-10.1	95,100
519	822	670	-15.7	45,100	604	783	1461	-16.7	14,400	682	1406	477	-6.1	59,100
520	632	165	-21.5	189,000	605	736	223	-18.0	125,300	683	1586	249	-3.4	109,800
521	1332	830	-7.1	37,300	606	629	273	-21.6	98,700	684	555	699	-24.8	43,500
522	603	1104	-22.6	26,600	607	1064	286	-10.8	94,000	685	1167	1313	-9.2	19,300
523	1190	309	-8.9	86,800	608	883	503	-14.5	56,700	686	1932	790	0.0	39,100
524	479	1226	-26.6	22,300	609	2012	610	>0.0	48,700	687	1545	619	-4.1	48,100
525	768	1066	-17.2	26,000	610	1255	903	-8.1	34,200	688	1456	764	-5.2	40,300
526	747	1016	-17.7	29,800	612	1103	391	-10.1	69,600	689	1011	953	-11.8	32,300
527	1170	231	-9.2	119,600	613	778	265	-16.9	102,000	690	1995	270	>0.0	100,200
528	1502	542	-4.6	53,400	614	824	518	-15.7	55,400	691	812	888	-16.0	34,900
530	1728	620	-2.0	48,000	615	1095	195	-10.3	149,100	692	1154	1461	-9.4	14,400
532	507	1011	-27.4	30,000	616	1759	478	-1.6	59,000	693	1993	819	>0.0	37,800
533	870	489	-14.7	57,900	617	994	372	-12.1	72,900	694	1628	656	-3.0	45,900
534	1347	1085	-6.9	27,300	618	751	374	-17.6	72,400	695	928	254	-13.6	107,000
535	1513	346	-4.5	77,800	619	1429	518	-5.7	55,300	696	1854	715	-0.6	42,700
536	308	654	<-35.0	46,000	620	1050	520	-11.1	55,200	697	1997	345	>0.0	78,000
538	1851	689	-0.7	44,100	621	923	1105	-13.7	26,600	698	957	563	-13.0	51,800
539	1463	982	-5.1	31,100	622	1462	622	-5.1	47,900	699	1540	730	-4.2	42,000
540	909	561	-13.9	52,000	623	759	225	-17.4	124,000	702	577	900	-23.8	34,400
541	625	289	-21.7	93,100	624	758	1038	-17.4	29,000	703	1610	562	-3.2	51,900
542	1164	198	-9.2	146,200	625	1438	606	-5.5	48,900	705	1278	571	-7.8	51,200
543	803	655	-16.2	45,900	626	1096	1089	-10.2	27,200	706	1841	704	-0.7	43,300
544	1259	1143	-8.0	25,200	627	942	548	-13.3	53,000	707	1018	1386	-11.7	16,900
545	856	1526	-15.0	12,200	628	809	621	-16.0	48,000	709	1074	1145	-10.7	25,100
546	803	1071	-16.2	27,800	629	899	979	-14.1	31,300	710	293	889	<-35.0	34,800
547	1162	274	-9.3	98,400	630	1135	1321	-9.6	19,100	712	720	412	-18.5	66,600
548	128	1321	<-35.0	19,000	631	979	615	-12.5	48,300	713	1386	841	-6.4	36,800
549	1355	1122	-6.8	25,900	632	1542	1076	-4.1	27,600	714	1328	263	-7.1	103,100
550	595	866	-23.0	35,800	633	1345	814	-6.9	38,000	715	698	433	-19.1	63,900
552	1369	494	-6.6	57,500	634	409	950	-32.2	32,400	716	701	481	-19.0	58,700
553	992	405	-12.2	67,600	635	1165	704	-9.2	43,300	717	1875	699	-0.5	43,600
555	1125	410	-9.8	66,900	636	774	604	-17.0	49,000	718	575	702	-23.9	43,400
556	705	975	-18.9	31,400	637	1263	524	-8.0	54,800	719	1216	204	-8.6	140,400
557	1477	1030	-4.9	29,300	638	952	411	-13.1	66,700	721	1069	464	-10.8	60,400
558	980	583	-12.5	50,400	639	1717	575	-2.1	51,000	722	1272	506	-7.9	56,400
559	700	1109	-19.1	26,400	640	994	292	-12.1	92,000	723	958	822	-13.0	37,700
560	1028	621	-11.5	48,000	641	165	1224	<-35.0	22,400	724	763	395	-17.3	69,100
562	898	794	-14.1	38,900	642	803	251	-16.2	108,900	725	720	916	-18.5	33,700
564	789	1446	-16.6	14,900	643	719	296	-18.5	90,700	726	1476	415	-4.9	66,200
565	777	766	-16.9	40,200	644	1100	294	-10.2	91,400	727	1846	473	-0.7	59,400
566	980	328	-12.5	81,900	645	534	1263	-26.1	21,000	728	510	783	-27.3	39,400
567	1519	611	-4.4	48,600	646	1153	1038	-9.4	29,000	729	1217	1126	-8.6	25,800
569	1212	661	-8.6	45,600	648	1246	204	-8.2	140,000	730	1858	724	-0.6	42,300
570	760	594	-17.4	49,700	649	14	1406	<-35.0	16,200	731	665	765	-20.2	40,300
571	618	956	-21.9	32,100	650	1713	1049	-2.1	28,600	733	1321	312	-7.2	85,900
573	1142	771	-9.6	40,000	651	1986	1183	>0.0	23,800	734	719	427	-18.5	64,600
574	532	787	-26.2	39,300	652	1378	816	-6.5	38,000	735	1101	473	-10.2	59,500
575	771	250	-17.1	109,200	653	1442	1165	-5.5	24,400	736	1359	569	-6.7	51,400
576	1068	534	-10.8	54,100	654	650	806	-20.8	38,400	738	696	220	-19.2	127,600
577	822	734	-15.7	41,800	655	1111	551	-10.0	52,700	739	687	409	-19.5	67,000
578	914	754	-13.8	40,800	656	1095	861	-10.3	36,000	740	1205	256	-8.7	106,200
579	1064	794	-10.8	38,900	657	1524	540	-4.4	53,600	741	995	563	-12.1	51,900
580	1524	714	-4.4	42,800	658	1777	860	-1.4	36,000	742	898	596	-14.1	49,500
581	1392	783	-6.3	39,400	659	391	584	-33.4	50,400	743	881	181	-14.5	165,900
582	962	686	-12.4	44,200	660	977	565	-12.5	51,700	744	1951	686	>0.0	44,200
584	1487	672	-4.8	45,000	661	658	166	-20.5	187,500	745	726	168	-18.3	183,600
585	758	731	-17.4	41,900	662	732	312	-18.1	86,100	746	999	643	-12.0	46,600
586	687	1152	-19.5	24,900	663	1787	567	-1.2	51,500	748	182	1503	<-35.0	13,000
587	930	523	-13.5	55,000	664	888	268	-14.4	100,900	749	2005	649	>0.0	46,300
588	1888	774	-0.4	39,900	665	889	775	-14.3	39,800	750	1448	575	-5.4	51,000
589	642	485	-21.1	58,300	666	715	221	-18.6	126,300	751	792	266	-16.5	101,900
590	1317	519	-7.3	55,300	667	781	227	-16.8	122,400	752	469	296	-28.9	90,600
591	65	1548	<-35.0	11,500	668	646	165	-21.0	189,100	754	664	254	-20.3	107,000
592	1014	614	-11.7	48,400	669	1116	353	-9.9	76,300	755	1195	184	-8.8	161,000
593	732	176	-18.1	172,300	670	1382	643	-6.4	46,600	756	1821	1113	-0.9	26,300
594	1627	478	-3.0	59,000	671	547	789	-25.3	39,200	757	909	246	-13.9	111,000
595	1009	1426	-11.8	15,500	673	984	746	-12.4	41,200	760	790	133	-16.5	264,900

MSN	X	Y	CPKdI	SDSMW	MSN	X	Y	CPKdI	SDSMW	MSN	X	Y	CPKdI	SDSMW
761	1399	733	-6.2	41,800	648	1863	271	-0.6	99,500	939	1197	827	-8.8	37,500
763	1416	1085	-5.9	27,300	649	1166	523	-9.2	54,900	941	1765	885	-1.5	35,000
764	2020	569	>0.0	51,400	650	1535	1024	-4.2	29,600	942	602	472	-22.7	56,600
765	651	475	-20.8	56,300	651	1035	826	-11.4	37,500	943	312	498	<-35.0	57,100
766	1052	1149	-11.1	25,000	652	834	542	-15.5	53,400	944	993	491	-12.1	57,700
767	1968	468	>0.0	56,900	655	499	220	-27.8	127,100	945	1300	269	-7.5	100,300
768	1330	685	-7.1	44,300	656	1063	194	-10.9	150,500	946	630	423	-21.6	65,100
769	1970	613	>0.0	48,500	657	887	890	-14.4	34,800	947	187	736	<-35.0	41,600
770	857	617	-15.0	48,200	658	1448	639	-5.4	46,900	948	1380	344	-6.5	78,200
771	1337	574	-7.0	31,500	659	706	311	-18.9	86,200	949	1766	665	-1.5	45,400
773	1576	502	-3.7	56,700	660	1070	1066	-10.7	28,000	950	1038	193	-11.3	151,000
775	969	824	-12.8	37,600	661	472	347	-28.8	77,600	951	860	152	-14.9	213,000
776	1438	708	-5.5	43,100	662	674	480	-19.9	58,800	952	957	701	-13.0	43,400
777	1539	458	-4.2	61,000	664	1307	499	-7.4	57,000	954	503	547	-27.6	53,000
778	850	434	-15.1	63,800	665	645	867	-21.0	34,900	955	1938	712	>0.0	42,900
779	700	411	-19.1	66,800	666	827	1004	-15.6	30,300	957	1010	816	-11.8	37,900
780	1052	1136	-11.1	25,500	668	685	494	-19.5	57,400	959	768	174	-17.2	174,900
784	1413	529	-6.0	54,400	669	1807	402	-1.0	68,000	960	596	419	-23.0	65,700
785	1364	885	-6.7	35,000	670	1323	783	-7.2	39,400	961	557	409	-24.8	67,100
786	1822	835	-0.9	37,100	671	1228	1031	-8.4	29,300	962	887	320	-14.4	83,900
787	893	392	-14.3	68,500	672	1904	346	-0.3	77,700	963	564	334	-24.5	80,500
790	616	882	-22.0	35,100	673	556	647	-24.8	46,400	964	969	1155	-12.8	24,800
791	451	1429	-29.8	15,400	674	1540	756	-4.2	40,700	965	671	255	-20.0	106,600
792	777	377	-16.9	72,000	675	1566	777	-3.8	39,700	966	1204	798	-8.7	38,700
793	1536	1543	-4.2	11,700	676	1198	351	-8.8	76,800	967	910	154	-13.9	210,300
794	1461	807	-5.1	36,300	677	1076	720	-10.6	42,500	968	609	1048	-22.3	28,700
796	388	546	-33.6	53,100	678	1161	1111	-9.3	26,400	969	1285	206	-7.7	138,900
797	1126	212	-9.8	133,700	679	647	757	-20.9	40,700	970	822	232	-15.8	119,300
798	933	437	-13.5	63,400	680	1756	594	-1.6	49,700	971	976	437	-12.6	63,400
799	1420	593	-5.9	49,800	681	1543	278	-4.1	97,100	972	403	567	-32.6	51,600
800	1759	279	-1.6	96,500	683	1432	890	-5.7	34,800	974	279	495	<-35.0	57,400
801	624	865	-21.7	35,800	684	922	689	-13.7	44,100	975	844	981	-15.3	31,200
802	898	547	-14.2	53,000	685	1103	414	-10.1	66,400	976	1124	295	-9.8	91,100
803	1775	1468	-1.4	14,200	686	1501	607	-4.6	48,900	977	994	664	-12.1	45,400
804	573	196	-24.0	146,400	687	798	1103	-16.3	26,600	978	1612	642	-3.2	46,700
805	203	494	<-35.0	57,400	688	636	634	-21.3	47,200	979	749	1141	-17.7	25,300
806	980	1039	-12.5	29,000	689	951	759	-13.1	40,600	980	1064	642	-10.8	46,700
807	902	308	-14.1	87,200	690	717	548	-18.6	52,900	981	1197	911	-8.8	33,900
808	625	827	-21.7	37,500	691	1123	229	-9.8	121,200	983	1762	1508	-1.6	12,800
809	1851	1015	-0.7	29,900	692	891	413	-14.3	66,400	984	1344	317	-6.9	84,700
810	440	573	-30.9	51,100	694	1245	234	-8.2	117,800	985	1024	1105	-11.5	26,600
811	1358	249	-6.8	109,700	695	1962	346	>0.0	77,700	987	739	1159	-17.9	24,600
812	851	383	-15.1	69,400	696	1322	626	-7.2	47,700	988	816	555	-15.9	52,400
813	745	1246	-17.8	21,600	697	420	570	-31.4	51,300	990	785	361	-16.7	74,900
814	2028	810	>0.0	38,200	698	662	428	-20.3	64,500	991	1159	317	-9.3	84,500
815	1086	645	-10.4	46,500	699	845	243	-15.3	113,000	992	1090	928	-10.4	33,300
816	629	313	-21.6	85,700	900	624	703	-21.7	43,400	993	1030	701	-11.5	43,400
817	1376	1177	-6.5	24,000	901	931	1094	-13.5	27,000	994	847	811	-15.2	38,200
818	1771	790	-1.4	39,100	903	799	229	-16.3	121,000	995	902	461	-14.1	60,700
819	1045	263	-11.2	103,100	904	765	520	-17.2	55,200	996	888	847	-14.4	36,600
820	984	362	-12.4	74,600	905	775	889	-17.0	34,800	997	1815	579	-0.9	50,700
821	1712	279	-2.2	96,700	907	888	824	-14.4	37,600	998	1205	504	-8.7	56,500
822	1256	205	-8.1	139,200	908	828	1303	-15.6	19,700	999	617	289	-22.0	93,100
823	1517	654	-4.4	46,000	910	681	1544	-19.7	11,700	1000	968	290	-12.8	92,700
824	1442	449	-5.5	62,000	911	1544	301	-4.1	89,100	1001	970	771	-12.7	40,000
825	1240	513	-8.3	55,800	913	1606	387	-3.3	70,400	1002	1736	478	-1.9	58,900
826	1309	1014	-7.4	29,900	914	1237	688	-8.3	44,100	1003	643	1184	-21.1	23,700
827	2012	708	>0.0	43,100	916	1442	749	-5.5	41,100	1006	822	487	-15.8	58,100
828	937	1405	-13.4	16,200	917	1260	367	-8.0	73,700	1007	875	279	-14.6	96,400
830	1342	756	-7.0	40,700	919	764	1541	-17.3	11,700	1009	291	644	<-35.0	46,600
831	562	826	-24.5	37,500	920	1133	1123	-9.7	25,900	1010	1386	745	-6.4	41,200
832	1073	1039	-10.7	29,000	921	1123	380	-9.8	71,500	1011	459	541	-29.4	53,500
833	481	820	-28.5	37,800	923	829	242	-15.6	113,200	1012	679	661	-19.7	45,600
834	501	581	-27.8	50,500	924	1131	318	-9.7	84,300	1013	1818	1128	-0.9	25,800
837	751	748	-17.6	41,100	925	1441	874	-5.5	35,400	1014	1032	634	-11.4	47,200
838	635	833	-21.3	37,200	926	679	219	-19.7	128,200	1015	1629	994	-3.0	30,700
839	1494	459	-4.7	60,900	927	1487	1191	-4.8	23,500	1016	1311	1134	-7.4	25,500
840	1952	301	>0.0	89,300	928	1082	775	-10.5	39,800	1017	1722	424	-2.0	65,000
841	1585	1080	-3.6	27,500	929	1231	816	-8.4	38,000	1018	1015	743	-11.7	41,300
842	571	1312	-24.1	19,400	931	1609	670	-3.3	45,100	1020	1574	1219	-3.7	22,500
843	1325	649	-7.2	46,300	932	810	900	-16.0	34,400	1021	781	484	-16.8	58,400
844	1727	301	-2.0	89,200	933	965	520	-12.8	55,100	1022	1129	83	-9.7	591,300
845	630	679	-21.5	44,600	934	947	462	-13.2	60,600	1023	812	317	-15.9	84,600
846	2016	905	>0.0	34,200	936	865	843	-14.8	36,800	1024	785	446	-16.7	62,400
847	673	1200	-19.9	23,200	937	1421	1056	-5.9	28,400	1025	1290	739	-7.7	41,500

MSN	X	Y	CPKpl	SDSMW	MSN	X	Y	CPKpl	SDSMW	MSN	X	Y	CPKpl	SDSMW
1026	405	532	-32.3	52,600	1153	921	1158	-13.7	24,700	1246	547	577	-25.3	50,800
1027	1298	848	-7.5	36,500	1154	1594	864	-3.5	35,900	1247	530	576	-26.3	50,900
1028	656	547	-15.0	53,000	1161	637	400	-21.3	66,400	1249	516	572	-27.0	51,200
1030	1284	226	-7.7	123,200	1162	623	397	-21.8	66,800	1250	973	536	-12.7	53,900
1031	686	622	-12.3	37,700	1163	665	397	-20.2	68,700	1251	607	532	-22.4	54,200
1032	1547	403	-4.1	67,900	1168	564	528	-24.4	54,500	1252	665	529	-20.2	54,400
1033	1381	551	-6.4	52,700	1170	552	529	-25.0	54,500	1253	899	766	-14.1	40,200
1034	1525	496	-4.3	57,200	1171	538	524	-25.9	54,800	1254	1311	746	-7.4	41,200
1035	1128	645	-9.7	46,500	1172	545	514	-25.5	55,700	1255	1300	761	-7.5	40,400
1036	1226	274	-8.5	96,300	1174	1099	522	-10.2	55,000	1257	1938	712	0.0	42,900
1039	1761	262	-1.6	103,600	1176	1304	586	-7.5	50,200	1258	1806	718	-1.0	42,600
1040	541	839	-25.7	36,900	1177	1366	539	-6.6	53,700	1259	1727	715	-2.0	42,700
1041	818	910	-15.8	34,000	1178	1608	702	-3.3	43,400	1260	1629	713	-3.0	42,800
1044	1036	485	-11.3	56,300	1179	1485	224	-4.8	124,900	1261	1555	717	-4.0	42,600
1045	1439	407	-5.5	67,300	1180	1459	224	-5.2	124,900	1262	1468	717	-5.0	42,600
1047	1540	250	-4.2	106,200	1181	1431	223	-5.7	125,100	1263	1413	722	-6.0	42,400
1048	1576	635	-3.7	47,100	1182	1407	223	-6.1	125,200	1264	1340	717	-7.0	42,600
1049	1089	411	-10.4	66,700	1183	1383	224	-6.4	124,700	1265	1263	717	-8.0	42,600
1050	949	1040	-13.2	26,900	1184	1454	182	-5.3	164,400	1266	1182	720	-9.0	42,500
1051	426	818	-31.1	37,800	1185	1422	183	-5.8	162,600	1267	1110	717	-10.0	42,600
1052	1583	1385	-3.6	16,900	1186	1394	182	-6.3	164,300	1268	1055	717	-11.0	42,600
1053	779	1092	-16.8	27,000	1189	1171	214	-9.2	131,800	1269	999	717	-12.0	42,600
1054	1613	620	-3.2	46,000	1190	1457	286	-5.2	94,200	1270	959	715	-13.0	42,700
1055	1380	377	-6.5	72,000	1191	686	1114	-19.5	26,200	1271	905	712	-14.0	42,900
1056	284	663	<-35.0	45,500	1192	265	893	<-35.0	34,700	1272	857	714	-15.0	42,800
1058	1261	746	-8.0	41,200	1193	403	1292	-32.6	20,000	1273	810	705	-16.0	43,300
1060	393	605	-33.3	49,000	1194	344	1275	<-35.0	20,600	1274	774	711	-17.0	42,900
1061	1817	645	-0.9	46,600	1195	505	1311	-27.6	19,400	1277	737	708	-18.0	43,100
1062	1245	746	-8.2	41,200	1196	572	1293	-24.1	20,000	1278	702	711	-19.0	42,900
1064	1258	792	-8.1	39,000	1197	639	1502	-21.2	13,000	1279	671	710	-20.0	43,000
1065	705	934	-18.9	33,000	1198	637	1402	-21.3	16,300	1280	645	710	-21.0	43,000
1066	1181	734	-9.0	41,800	1199	614	1407	-22.1	16,200	1281	617	707	-22.0	43,100
1067	529	658	-26.3	45,800	1200	637	1431	-21.3	15,400	1282	595	704	-23.0	43,300
1068	508	696	-27.4	43,700	1201	1095	1394	-10.3	16,600	1283	573	700	-24.0	43,500
1069	1898	604	-0.3	49,100	1202	1719	1545	-2.1	11,600	1284	552	695	-25.0	43,700
1071	873	609	-14.7	48,700	1203	791	668	-16.5	45,200	1285	536	694	-26.0	43,800
1073	1768	1128	-1.5	25,800	1204	964	1021	-12.9	29,700	1286	515	687	-27.0	44,200
1075	836	773	-15.4	39,900	1205	313	195	<-35.0	148,700	1287	496	683	-28.0	44,400
1076	1863	861	-0.6	36,000	1208	306	194	<-35.0	149,800	1288	467	669	-29.0	45,200
1078	826	566	-15.7	51,600	1209	320	197	<-35.0	147,400	1289	447	667	-30.9	45,300
1081	671	483	-12.7	56,500	1210	326	197	<-35.0	146,600	1290	427	655	-31.0	45,900
1083	1697	202	-2.3	142,300	1211	394	294	-33.2	91,400	1291	412	655	-32.0	45,900
1085	1157	794	-9.4	36,900	1212	402	294	-32.7	91,200	1292	397	652	-33.0	46,100
1090	620	910	-21.9	34,000	1214	386	294	-33.7	91,400	1293	381	654	-34.0	46,000
1092	1867	597	-0.5	49,500	1215	641	329	-21.2	81,600	1294	365	653	-35.0	46,100
1093	2019	894	>0.0	34,600	1216	660	329	-20.4	81,600	1295	348	653	<-35.0	46,100
1094	1546	538	-4.1	53,700	1217	914	266	-13.8	101,800					
1095	1545	477	-4.1	59,100	1218	873	245	-14.7	112,000					
1098	61	935	<-35.0	33,000	1219	970	372	-12.7	72,900					
1099	1954	237	>0.0	116,000	1220	1021	298	-11.6	90,100					
1101	588	1048	-23.3	28,600	1221	1392	205	-6.3	139,500					
1102	1050	667	-11.1	45,200	1222	1354	203	-6.8	141,800					
1103	457	797	-29.5	38,800	1223	1362	205	-6.7	139,500					
1105	1884	532	-0.4	54,200	1224	673	540	-19.9	53,600					
1106	1714	649	-2.1	46,300	1225	614	542	-22.1	53,400					
1107	1717	546	-2.1	53,100	1226	603	539	-22.6	53,600					
1108	1976	722	>0.0	42,400	1227	696	623	-19.2	47,800					
1111	547	1066	-25.3	28,000	1228	707	628	-18.9	47,500					
1112	1348	621	-6.9	48,000	1229	475	447	-28.7	62,300					
1115	1385	762	-6.4	40,400	1230	466	1282	-29.0	20,400					
1116	1078	816	-10.6	38,000	1231	759	1461	-17.4	14,400					
1117	975	787	-12.6	39,300	1232	1324	1170	-7.2	24,200					
1118	1202	933	-8.7	33,100	1233	1583	1005	-3.6	30,300					
1119	1022	1076	-11.6	27,600	1234	1865	809	-0.6	38,200					
1120	1905	616	-0.3	48,300	1235	1812	817	-1.0	37,900					
1121	1512	1301	-4.5	19,700	1236	1411	703	-6.0	43,400					
1122	1114	677	-9.9	44,700	1237	1392	682	-6.3	44,500					
1123	1464	452	-5.1	61,700	1238	794	410	-16.4	66,900					
1125	1048	857	-11.1	36,200	1239	769	407	-17.1	67,300					
1126	1122	802	-9.8	38,600	1240	740	406	-17.9	67,500					
1128	1722	892	-2.1	34,700	1241	743	511	-17.8	55,900					
1133	1098	825	-10.2	37,500	1242	713	510	-18.7	56,000					
1139	1830	569	-0.8	51,400	1243	682	509	-19.6	56,100					
1147	764	1182	-17.3	23,800	1244	663	504	-20.3	56,500					
1148	1968	724	>0.0	42,300	1245	565	582	-24.4	50,500					

Table 2. Table of some identified proteins

POP name	Protein name	MSN's	Basis for identification
IDS:3_ALPHA_HDDH	3- α -hydroxysteroid-dihydroxyl-dehydrogenase, an enzyme of steroid metabolism	137, 159	Pure protein and antibody provided by Dr. T.M. Penning, Department of Pharmacology, School of Medicine, University of Pennsylvania.
IDS:ACTIN_BETA	β cellular actin, a cytoskeletal protein	38	Homologous position with respect to other mammalian systems
IDS:ACTIN_GAMMA	γ cellular actin, a cytoskeletal protein	68	Homologous position with respect to other mammalian systems
IDS:ALBUMIN	Serum albumin, mature form.	21, 28, 33	Predominance in rat plasma
IDS:APO_A-I	Apo A-I plasma lipoprotein, mature form (tentative)	236, 463	Presence in rat plasma, regulation by some lipid-lowering drugs
IDS:CALMODULIN	Calmodulin, an acidic cytosolic calcium-binding protein	123, 649	Homologous position with respect to other mammalian systems
IDS:CATALASE	Catalase (peroxisomal)	54, 61, 106	Presence in purified peroxisomes, similarity in position to mouse catalase
IDS:CPKSPOTS	Spots contributed by the CPK charge standards (not rat liver proteins)	1257 - 1295	
IDS:CPS	Carbamoyl phosphate synthase	114, 157, 167, 174, 1184, 1185, 1186, 1222	Pure protein provided by Dr. Margaret Marshall, Department of Pharmacology, Medical School, University of Wisconsin - Madison.
IDS:CYTOCHROME_B5	Cytochrome b5	87, 477	Pure protein provided by Dr. Andrew Parkinson, Department of Pharmacology, Toxicology and Therapeutics, University of Kansas Medical Center
IDS:FABP-L	Liver fatty-acid binding protein	227	Pure protein provided by Dr. Nathan Bass, Department of Medicine, University of California School of Medicine, San Francisco
IDS:HMG-COA_SYNTHASE	Cytosolic HMG-CoA Synthase	133, 144, 235, 413	Antibody provided by Dr. Michael Greenspan, Merck Sharp & Dohme Research Laboratories, Rahway, NJ
IDS:LAMIN_B	Lamin B, a nuclear protein	415, 734	Homologous position with respect to other mammalian systems
IDS:MITCON:1	Mitcon:1 (F1 ATPase β subunit), a mitochondrial inner membrane	17, 49, 71, 340, 1245, 1246, 1247, 1249	Homologous position with respect to other mammalian systems, presence in mitochondria
IDS:MITCON:2	Mitcon:2, a mitochondrial matrix stress protein equivalent to E.	15, 25, 110, 1241, 1242, 1243, 1244	Homologous position with respect to other mammalian systems, presence in mitochondria
IDS:MITCON:3	Mitcon:3, a mitochondrial matrix stress protein, likely analog of	18, 35, 226, 600, 1238, 1239, 1240	Homologous position with respect to other mammalian systems, presence in mitochondria
IDS:NADPH_P450_RED	NADPH cytochrome P-450 reductase, frequently co-induced with P-450's	175, 251, 812	Pure protein provided by Dr. Andrew Parkinson, Department of Pharmacology, Toxicology and Therapeutics, University of Kansas Medical Center
IDS:PD1	Protein disulphide isomerase 1	168, 1170, 1171, 1172	Sequence information obtained by R.M. Van Frank, Lilly Research Laboratories, Indianapolis
IDS:PLASMA_PROTEINS	Rat plasma proteins observed in liver	21, 28, 33, 44, 72, 102, 115, 197, 236, 246, 248, 257, 293, 332, 347, 364, 369, 419, 432, 463, 468, 518, 562, 605, 623, 666, 667, 725, 738, 790, 865, 903, 926, 97, 93	Plasma coelectrophoresis studies
IDS:PRO-ALBUMIN	Serum albumin precursor	179, 1180, 1181, 1182, 1183	Relative position to mature albumin, presence in microsomes
IDS:PYRCARBOX	Pyruvate carboxylase	135	Pavlica, R.J., et al., BBA (1990) 1022 115-125.
IDS:SOD	Superoxide dismutase	56, 132, 1224, 1252	Sequence information obtained by R.M. Van Frank, Lilly Research Laboratories, Indianapolis
IDS:TUBULIN_ALPHA	α tubulin, a cytoskeletal protein	50, 1225, 1226, 1251	Homologous position with respect to other mammalian systems
IDS:TUBULIN_BETA	β tubulin, a cytoskeletal protein		Homologous position with respect to other mammalian systems

Table 3. Computed pI's of two sets of carbamylated protein standards: Rabbit muscle CPK and human hemoglobin (Hb)

	Protein Name	PIR Name	#ASP 3.9	#GLU 4.1	#HIS 6.0	#LYS 10.8	#ARG 12.5	NH2- 7.0	Calc pI	Real CPK
0	Rabbit muscle CPK	KIRBCM	28	27	17	34	18	1	6.84	0.0
-1			28	27	17	33	18	1	6.67	-1
-2			28	27	17	32	18	1	6.54	-2
-3			28	27	17	31	18	1	6.42	-3
-4			28	27	17	30	18	1	6.31	-4
-5			28	27	17	29	18	1	6.21	-5
-6			28	27	17	28	18	1	6.12	-6
-7			28	27	17	27	18	1	6.03	-7
-8			28	27	17	26	18	1	5.94	-8
-9			28	27	17	25	18	1	5.85	-9
-10			28	27	17	24	18	1	5.76	-10
-11			28	27	17	23	18	1	5.67	-11
-12			28	27	17	22	18	1	5.58	-12
-13			28	27	17	21	18	1	5.48	-13
-14			28	27	17	20	18	1	5.39	-14
-15			28	27	17	19	18	1	5.29	-15
-16			28	27	17	18	18	1	5.20	-16
-17			28	27	17	17	18	1	5.12	-17
-18			28	27	17	16	18	1	5.04	-18
-19			28	27	17	15	18	1	4.96	-19
-20			28	27	17	14	18	1	4.89	-20
-21			28	27	17	13	18	1	4.83	-21
-22			28	27	17	12	18	1	4.77	-22
-23			28	27	17	11	18	1	4.71	-23
-24			28	27	17	10	18	1	4.66	-24
-25			28	27	17	9	18	1	4.61	-25
-26			28	27	17	8	18	1	4.56	-26
-27			28	27	17	7	18	1	4.52	-27
-28			28	27	17	6	18	1	4.48	-28
-29			28	27	17	5	18	1	4.44	-29
-30			28	27	17	4	18	1	4.40	-30
-31			28	27	17	3	18	1	4.36	-31
-32			28	27	17	2	18	1	4.32	-32
-33			28	27	17	1	18	1	4.29	-33
-34			28	27	17	0	18	1	4.25	-34
-35			28	27	17	0	18	0	4.22	-35
0	Hb-beta, human	HBHU	7	8	9	11	3	1	7.18	
-1			7	8	9	10	3	1	6.79	
-2			7	8	9	9	3	1	6.53	-1.8
-3			7	8	9	8	3	1	6.32	-3.2
-4			7	8	9	7	3	1	6.13	-5.3
-5			7	8	9	6	3	1	5.96	-7.2
-6			7	8	9	5	3	1	5.78	-10.0
-7			7	8	9	4	3	1	5.59	-12.3
-8			7	8	9	3	3	1	5.37	-15.5
-9			7	8	9	2	3	1	5.14	-18.0
-10			7	8	9	1	3	1	4.91	-21.0
-11			7	8	9	0	3	1	4.71	-25.5
-12			7	8	9	0	3	0	4.54	-27.2

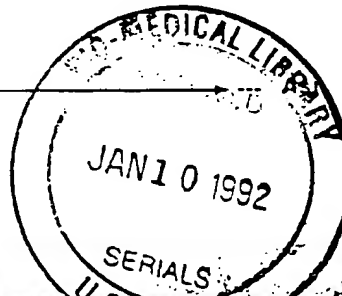
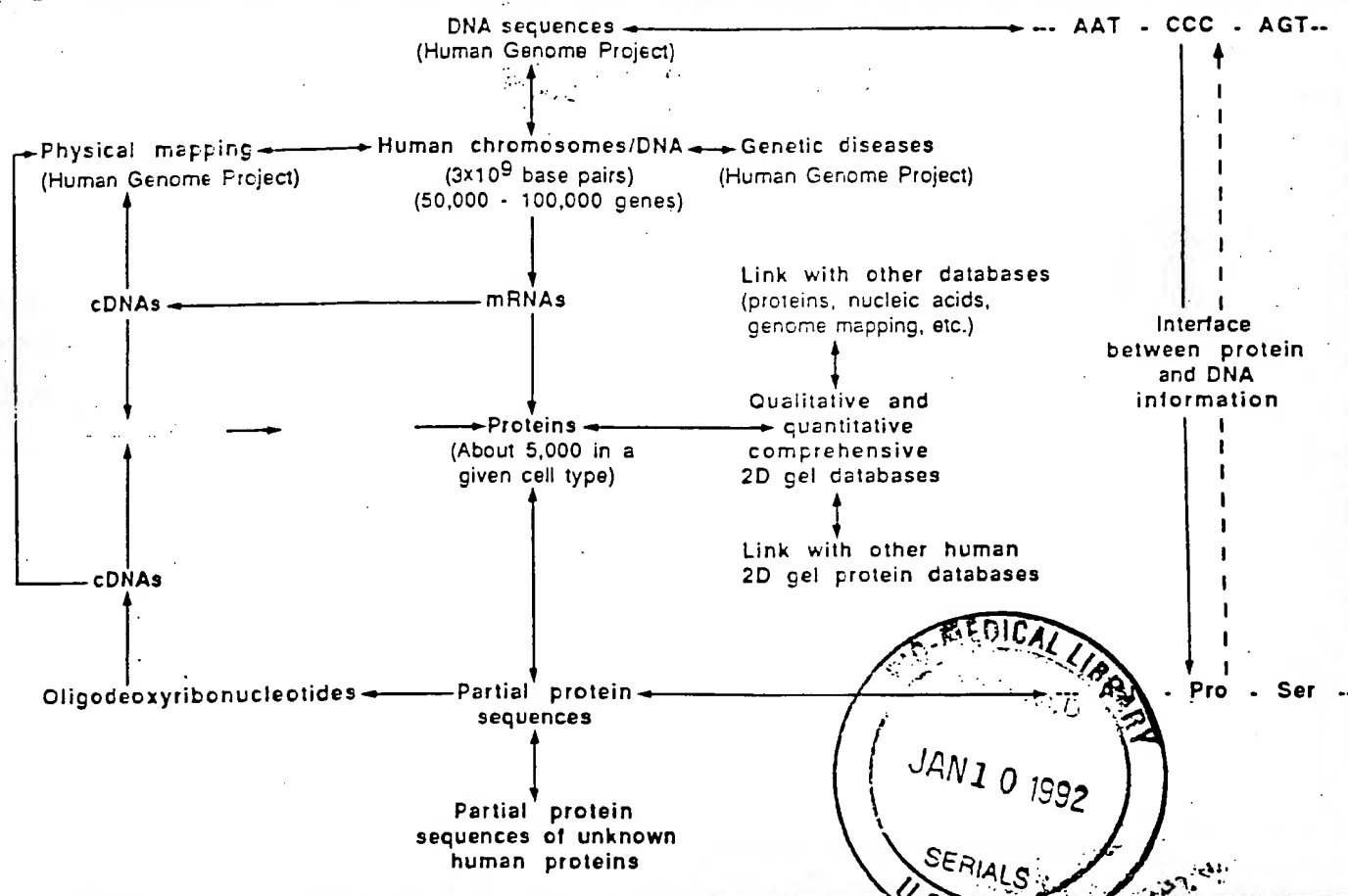
Table 4. Computed pI's of some known proteins related to measured CPK pI's

	Protein Name	PIR Name	#ASP 3.9	#GLU 4.1	#HIS 6.0	#LYS 10.8	#ARG 12.5	Calc pI	Real CPK
0	Creatine phospho kinase (CPK), rabbit muscle	KIRBCM	28	27	17	34	18	6.84	0.0
1	Fatty acid-binding protein, rat hepatic	FZRTL	5	13	2	16	2	7.83	-3.0
2	b2-microglobulin, human	MGHUB2	7	8	4	8	5	6.09	-5.0
3	Carbamoyl-phosphate synthase, rat	SYRTCA	72	96	28	95	56	5.97	-5.5
4	Proalbumin (serum albumin precursor), rat	ABRTS	32	57	15	53	27	5.98	-6.2
5	Serum albumin, rat	ABRTS	32	57	15	53	24	5.71	-9.0
6	Superoxid dismutase (Cu-Zn, SOD), rat	A26810	8	11	10	9	4	5.91	-9.2
7	Phospholipase C, phosphoinositide-specific (?), rat	A26607	34	42	9	49	21	5.92	-9.2
8	Albumin, human	ABHUS	36	61	16	60	24	5.70	-11.9
9	Apo A-I lipoprotein, rat	A24700	18	24	6	23	12	5.32	-13.7
10	proApo A-I lipoprotein, human	LPHUA1	16	30	6	21	17	5.35	-14.3
11	NADPH cytochrome P-450 reductase, rat	RDRT04	41	60	21	38	36	5.07	-15.6
12	Retinol binding protein, human	VAHU	18	10	2	10	14	5.04	-16.9
13	Actin beta, rat	ATRTC	23	26	9	19	18	5.06	-17.2
14	Actin gamma, rat	ATRTC	20	29	9	19	18	5.07	-16.8
15	Apo A-I lipoprotein, human	LPHUA1	16	30	5	21	16	5.10	-17.5
16	Apo A-IV lipoprotein, human	LPHUA4	20	49	8	28	24	4.88	-19.7
17	Tubulin alpha, rat	UBRTA	27	37	13	19	21	4.66	-19.8
18	F1ATPase beta, bovine	PWBOB	25	36	9	22	22	4.80	-21.0
19	Tubulin beta, pig	UEPGB	26	36	10	15	22	4.49	-22.5
20	Protein disulphide isomerase (PDI), rat hepatic	ISRTSS	43	51	11	51	9	4.07	-25.0
21	Cytochrome b5, rat	CBRT5	10	15	6	10	4	4.59	-26.0
22	Apo C-II lipoprotein, human	LPHUC2	4	7	0	6	1	4.44	-30.5
Amino acid pI assumed in calculation:			3.9	4.1	6.0	10.8	12.5		

ELECTROPHORESIS

An International Journal

11 91



TWO-DIMENSIONAL GEL PROTEIN DATABASES

Editor J. E. Celis



ELECTROPHORESIS

An International Journal

Indexed in: BIOSIS
Current Contents, MEDLINE
ISSN 0173-0833
ELCTDN 12 (11) 763-996 (1991)

TWO-DIMENSIONAL GEL PROTEIN DATABASES

Editor: J. E. Celis

Editorial

- J. E. Celis, H. Leffers,
H. H. Rasmussen, P. Madsen,
B. Honoré, B. Gesser, K. Dejgaard,
E. Olsen, G. P. Ratz, J. B. Lauridsen,
B. Basse, A. H. Andersen,
E. Walbum, B. Brandstrup, A. Celis,
M. Puype, J. Van Damme and
J. Vandekerckhove
- 765 The master two-dimensional gel database of human AMA cell proteins: Towards linking protein and genome sequence and mapping information (Update 1991)
- J. E. Celis, F. Madsen,
H. H. Rasmussen, H. Leffers,
B. Honoré, B. Gesser, K. Dejgaard,
E. Olsen, N. Magnusson, J. Kiil,
A. Celis, J. B. Lauridsen, B. Basse,
G. P. Ratz, A. H. Andersen,
E. Walbum, B. Brandstrup,
P. S. Pedersen, N. J. Brandt,
M. Puype, J. Van Damme and
J. Vandekerckhove
- 802 A comprehensive two-dimensional gel protein database of noncultured unfractionated normal human epidermal keratinocytes: Towards an integrated approach to the study of cell proliferation, differentiation and skin diseases
- H. H. Rasmussen, J. Van Damme,
M. Puype, B. Gesser, J. E. Celis and
J. Vandekerckhove
- 873 Microsequencing of proteins recorded in human two-dimensional gel protein databases
- N. L. Anderson and N. G. Anderson
- 883 A two-dimensional gel database of human plasma proteins
- N. L. Anderson, R. Esquer-Blasco,
J.-P. Hofmann and N. G. Anderson
- 907 A two-dimensional gel database of rat liver proteins useful in gene regulation and drug effects studies
- P. J. Wirth, L.-di Luo, Y. Fujimoto,
H. C. Bisgaard and A. D. Olson
- 931 The rat liver epithelial (RLE) cell protein database
- R. A. VanBogelen and
F. C. Neidhardt
- 955 The gene-protein database of *Escherichia coli*: Edition 4
- 995 Miscellaneous

For submission of papers, see Instructions to Authors (last page of this issue)

High Specific Activity Chemiluminescent and Fluorescent Markers: their Potential Application to High Sensitivity and 'Multi-analyte' Immunoassays

Roger Ekins*, Frederick Chu and Jacob Micallef

Department of Molecular Endocrinology, University College and Middlesex School of Medicine, University of London, Mortimer Street, London W1N 8AA, UK

The sensitivities of immunoassays relying on conventional radioisotopic labels (i.e. radioimmunoassay (RIA) and immunoradiometric assay (IRMA)) permit the measurement of analyte concentrations above ca 10^7 molecules/ml. This limitation primarily derives, in the case of 'competitive' or 'limited reagent' assays, from the 'manipulation errors arising in the system combined with the physicochemical characteristics of the particular antibody used; however, in the case of 'non-competitive' systems, the specific activity of the label may play a more important constraining role. It is theoretically demonstrable that the development of assay techniques yielding detection limits significantly lower than 10^7 molecules/ml depends on:

- (1) the adoption of 'non-competitive' assays designs;
- (2) the use of labels of higher specific activity than radioisotopes;
- (3) highly efficient discrimination between the products of the immunological reactions involved.

Chemiluminescent and fluorescent substances are capable of yielding higher specific activities than commonly used radioisotopes when used as direct reagent labels in this context, and both thus provide a basis for the development of 'ultra-sensitive', non-competitive, immunoassay methodologies. Enzymes catalysing chemiluminescent reactions or yielding fluorescent reaction products can likewise be used as labels yielding high effective specific activities and hence enhanced assay sensitivities.

A particular advantage of fluorescent labels (albeit one not necessarily confined to them) lies in the possibility they offer of revealing immunological reactions localized in 'microspots' distributed on an inert solid support. This opens the way to the development of an entirely new generation of 'ambient analyte' microspot immunoassays permitting the simultaneous measurement of tens or even hundreds of different analytes in the same small sample, using (for example) laser scanning techniques. Early experience suggests that microspot assays with sensitivities surpassing that of isotopically based methodologies can readily be developed.

Keywords: Ultrasensitive immunoassay; fluorescent microspot immunoassay; confocal microscopy

* Author for correspondence.

INTRODUCTION

Immunoassay methods relying on radioisotopic labels have played a major role in medicine and other biologically related fields (agriculture, veterinary science, the food and pharmaceutical industries, etc.) during the past two decades. Their importance has derived from the exploitation both of the 'structural specificity' characterizing antibody-antigen reactions and the 'detectability' of isotopically-labelled reagents, the latter permitting observation of the binding reactions between exceedingly small concentrations of the key reactants involved. The combination of these features has endowed radioimmunoassay methods with unique specificity and sensitivity characteristics, and accounts for their ubiquitous use throughout modern medicine and biology. However, in the past few years, interest has increasingly focused on so-called 'alternative', non-radioisotopic, immunoassay methods; such techniques are based on essentially identical analytical principles but differ in the markers used to label the particular immunoreactant (antibody or analyte) whose distribution between bound and free moieties (following the basic analytical reaction) constitutes the assay 'response'. The reasons for this interest may be grouped under four headings:

- (1) Environmental; logistic; economic; practicality and convenience, etc. (i.e. 'non-scientific').
- (2) The attainment of higher sensitivity.
- (3) The development of 'immunosensors' and 'immunoprobes'.
- (4) The development of 'multi-analyte' assay systems.

Our own reasons for developing non-isotopic techniques fall principally under headings (2) and (4), and this presentation will centre primarily on the concepts which underlie our immunoassay development strategy in these areas.

THE ATTAINMENT OF 'ULTRA-HIGH' IMMUNOASSAY SENSITIVITY

Though, as indicated above, the sensitivity of radioisotopically based immunoassay methods has constituted one of the principal foundations of their widespread use over the past 25 years, a

fundamental reason for their replacement stems, paradoxically, from the current requirement to develop microanalytical techniques which are superior to them in this particular respect. Radioisotopic methods are, in practice, limited to the measurement of analyte concentrations above about 10^8 – 10^9 molecules/ml (i.e. approx 0.15–1.5 pmol/l) (Dakubu *et al.*, 1984). However, in certain fields (e.g. virology, tumour detection) there is a particular need to detect or measure molecular concentrations below this level. The factors which determine immunoassay sensitivity have been extensively discussed (Ekins *et al.*, 1968, 1970a; Ekins, 1978; Jackson *et al.*, 1983; Dakubu *et al.*, 1984; Ekins, 1985). Nevertheless, some of the underlying concepts are still frequently misunderstood and merit brief discussion in the present context.

The concept of sensitivity

One major source of past confusion has been disagreement regarding the concept of 'sensitivity' itself, many authors equating assay sensitivity with the slope of the dose-response curve (Yalow and Berson, 1970a, b; Berson and Yalow, 1973; see also Ekins *et al.*, 1970b, Tait, 1970). It is now widely agreed that the notion that a steeper dose-response curve implies greater sensitivity is erroneous. The invalidity of this belief is clearly revealed by the fact that the relative magnitudes of the responses yielded by two assay systems is dependent on the particular variable which is chosen to represent the response (see Fig. 1(a)) (Ekins, 1976). For this and other reasons, it has long been recognized that the 'sensitivity' of an assay can only be satisfactorily represented by its lower limit of detection (Fig. 1(b)), and this concept is now embodied in all internationally agreed definitions of the term. An essentially identical definition is as the precision (i.e. standard deviation) of measurement of zero dose, since this quantity determines the least quantity distinguishable from zero and hence the assay detection limit. The sensitivity of an assay is thus represented by the zero-dose intercept of the 'precision profile' (Fig. 2(a)) when the latter is expressed in terms of standard deviation rather than of coefficient of variation (Ekins, 1983a). In short, the more sensitive of two assays is the one yielding greater precision of the zero dose estimate (Fig. 2(b)).

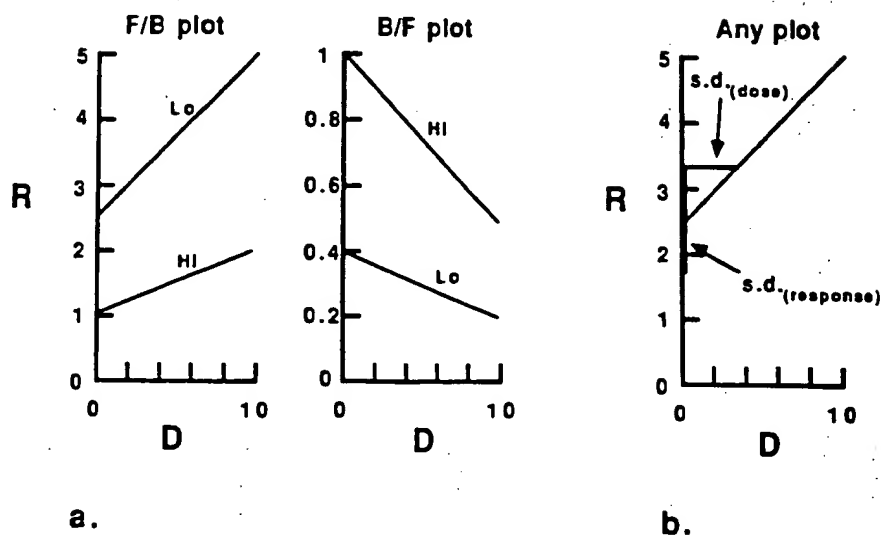


Figure 1. (a) Diagrammatic representation of conventional RIA dose-response curves for systems using high (hi) and low (lo) antibody concentrations plotted in terms of free-bound (F/B) and bound/free (B/F) labelled antigen. Note that the use of a lower amount of antibody yields a dose-response curve of greater slope in the F/B plot, but of lower slope in the B/F plot. It is impossible to decide, on the basis of the data shown in this figure, which concentration of antibody yields the assay system of higher sensitivity. (b) The sensitivity of an assay is essentially represented by the minimum detectable dose, i.e. the SD of the dose measurement ($SD_{(dose)}$) at zero dose. This is given by the SD of the response ($SD_{(response)}$) divided by the dose-response curve slope at zero dose (i.e. $(SD_{(response)}) \times dD/dR)_0$). This quantity is unaffected by the choice of the coordinate frame used to plot the dose-response curve. (Note: it is common to multiply $(SD_{(dose)})_0$ by an arbitrary factor to increase the confidence level attaching to the minimum detectable dose estimate, though, since no agreement exists regarding the value of this factor, this unnecessary step merely adds to confusion when the relative sensitivities of two assay procedures are compared.)

'Competitive' and 'non-competitive' ('limited reagent' and 'excess reagent') assays

A second important misconception in this area is the notion that immunoassays relying on the use of *labelled antibodies* (e.g. immunoradiometric assays, IRMA) are *ipso facto* more sensitive than

those which rely on the use of *labelled 'analyte'* (e.g. radioimmunoassays, RIA); furthermore the grounds originally advanced for the claimed superiority of labelled antibody methods (Miles and Hales, 1968) were partially based on false concepts of sensitivity, and thus failed to identify the *true* reasons why certain assay designs are

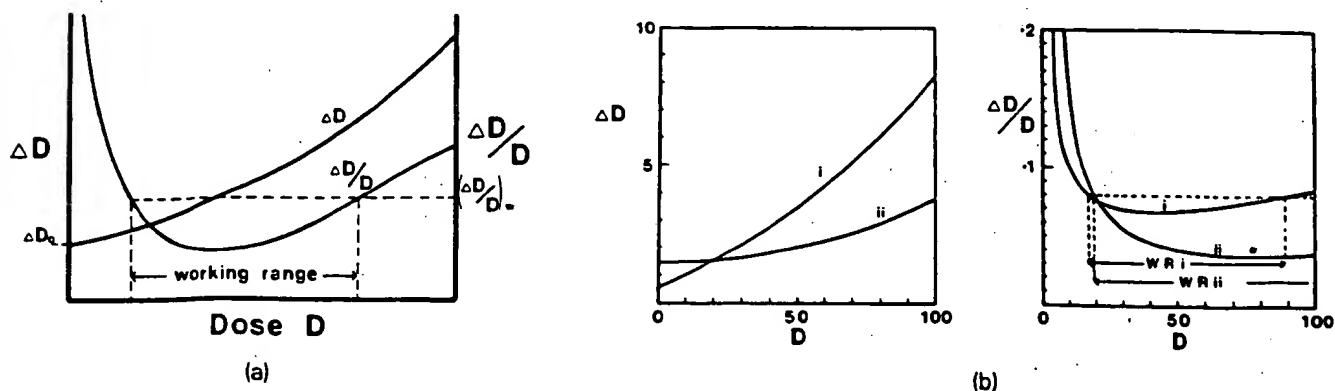


Figure 2. (a) The 'precision profile' of an assay portrays the error in the dose measurement as a function of dose. The error may be represented, *inter alia*, by the absolute error (ΔD ; e.g. SD of D) or the relative error ($\Delta D/D$; e.g. CV of D). $(\Delta D)_0$, the error in the measurement of zero dose, represents the sensitivity of the assay. The working range may be defined as the range of dose values within which $\Delta D/D$ is less than an 'acceptable' value set by the investigator. (b) The more sensitive of the two assays (assay I) intercepts the ΔD axis at a lower value. However, assay II is more precise at higher values of dose, and has a wider working range.

potentially capable of yielding far higher sensitivity than others. This issue likewise merits clarification.

The purely pragmatic sub-classification of immunoassays into labelled antibody and labelled analyte methods diverts attention from a more fundamental divide in immunoassay methodology, which relates to the optimal concentration of antibody required in an assay system to maximize its sensitivity. In certain assay designs (which may be termed 'limited reagent' or 'competitive') the optimal concentration tends to zero; conversely in others (which may be termed 'excess reagent' or 'non-competitive') the concentration tends to infinity. It should be particularly emphasized that the optimal antibody concentration is essentially governed, not only by the physicochemical characteristics of the antibody-analyte binding reaction, but also by the errors incurred in measurement of the assay response. Were an assay system to be totally error-free, *no* antibody concentration would be optimal, and the distinction between competitive and non-competitive methodologies would thus not arise.

Though it is inappropriate in this presentation to discuss in detail the statistical and physicochemical theory underlying this fundamental divergence in immunoassay design (see Ekings *et al.*, 1968, 1970a; Jackson *et al.*, 1983), the reason for it can perhaps be more readily understood if the basic principles of immunoassay are portrayed in a somewhat different way from that in which they are usually presented. All immunoassays essentially depend upon measurement of the 'fractional occupancy' by analyte of antibody binding sites following reaction of analyte with antibody (see Fig. 3(a)). Those techniques which implicitly rely on measurement of residual, *unoccupied*, binding sites optimally necessitate the use of concentrations of antibody tending to zero, and may be termed 'competitive', conversely those in which *occupied* sites are directly measured necessitate use of high antibody concentrations and are termed 'non-competitive' (Fig. 3(b)). This emphasizes that the differences in assay design characterizing so-called competitive and non-competitive methods are essentially unrelated to which component (if any) of the reaction system is labelled. Indeed immunoassays in which *no label of any kind is involved* can, on identical grounds, be subdivided into those of 'limited reagent' (or 'competitive') and 'excess reagent' (or 'non-competitive') design. Thus the

distinction between these two forms of immunoassay simply reflects differences in the way that fractional antibody occupancy is determined, and the fact that it is generally undesirable—for reasons of accuracy—to measure a *small* quantity by estimating the difference between two *large* quantities. When an immunoassay relies on the measurement of unoccupied antibody binding sites, the total amount of antibody used in the system must be small to minimize error in the resulting (indirect) estimate of occupied sites.

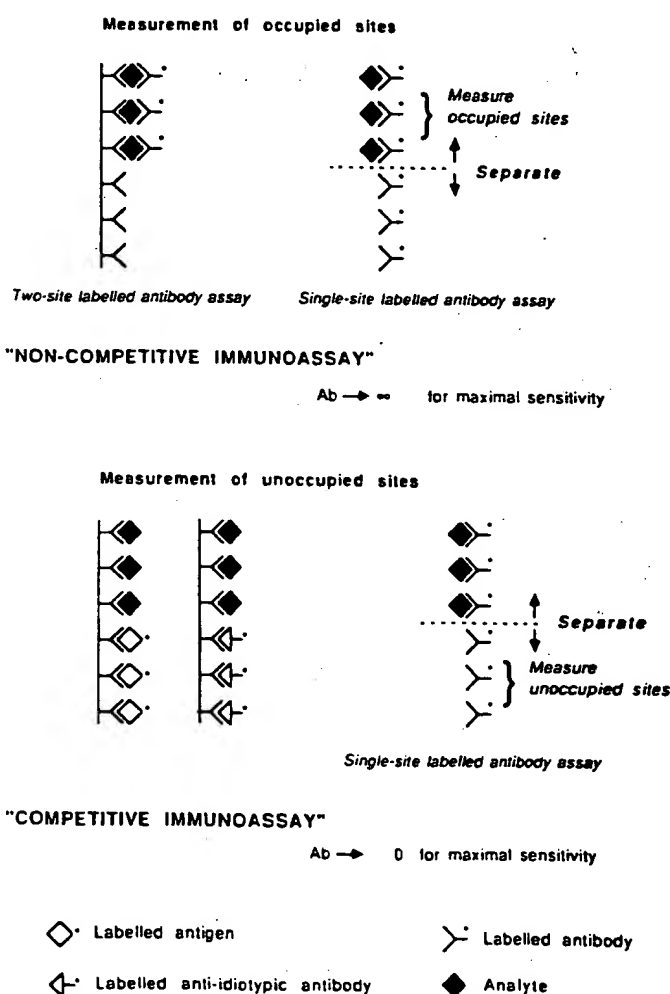


Figure 3. The distinction between 'non-competitive' (above) and 'competitive' immunoassays (below) reflects how antibody binding-site occupancy is measured. Labelled antibody methods are 'non-competitive' if occupied sites of the (labelled) antibody are measured, but are 'competitive' (below right) when *unoccupied* sites are measured. Labelled antigen (below left) or labelled anti-idiotypic antibody methods (below centre) rely on measurement of sites *unoccupied* by analyte, and are therefore invariably of 'competitive' design.

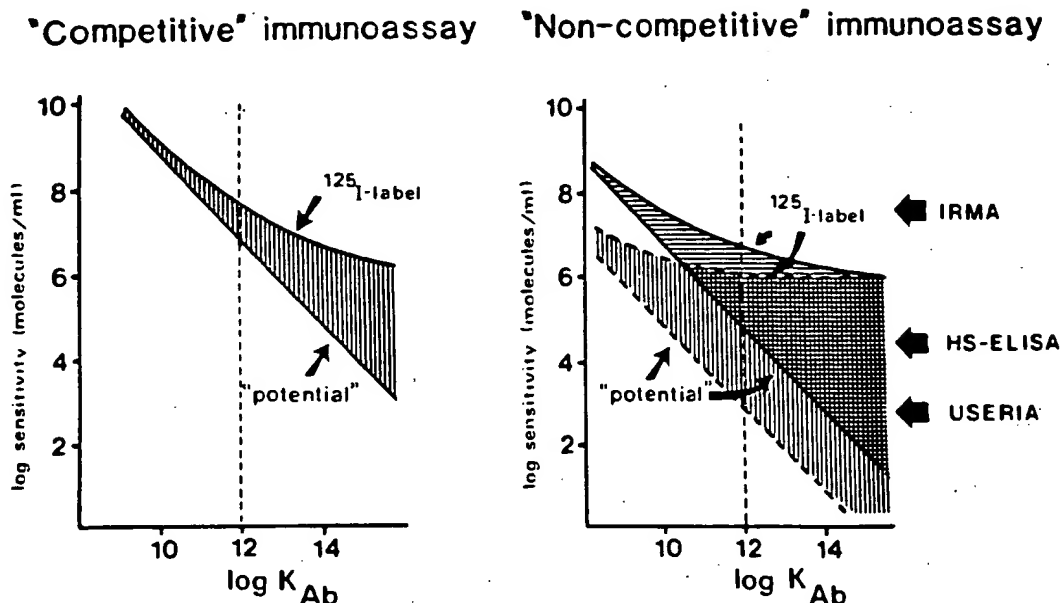


Figure 4. Curves showing the theoretically predicted relationship between antibody affinity and the sensitivities achievable using 'competitive' and 'non-competitive' assay strategies. The 'potential' sensitivity curves assume the use of infinite specific activity labels; the sensitivities achievable using ^{125}I -labelled antigen or antibody are also shown. Shaded areas indicate the sensitivity loss due to errors in measurement of the label. Curves relating to 'competitive' assays assume a 1% error in measurement of the response variable arising from 'experimental' errors (i.e. errors other than those inherent in label measurement *per se*). Non-competitive curves assume 'non-specific binding' of labelled antibody of 0.01% and 1% (lower and upper curves) respectively. Arrows indicate sensitivities claimed for typical non-competitive immunoassay methodologies.

Conversely, when occupied sites are measured *directly*, this particular constraint does not arise; indeed, considerable advantage often derives from using relatively large amounts of antibody in the system.

Sensitivity of 'competitive' and 'non-competitive' immunoassays

Competitive and non-competitive immunoassays differ significantly in many of their performance characteristics in consequence of the differences in optimal antibody concentration on which they rely. Most particularly they differ in their potential sensitivities. Figure 4. portrays the sensitivities predicted theoretically as a function of antibody binding affinity, making realistic assumptions regarding the experimental errors incurred in reagent manipulation, 'non-specific' binding of labelled antibody, etc., and assuming the use of optimal reagent concentrations (Ekins, 1985). Amongst other concepts illustrated in the figure is the much greater assay sensitivity *potentially* attainable (using an antibody of given affinity) by adoption of a non-competitive approach. In short, whereas the maximal sensitiv-

ity realistically achievable using a competitive design is in the order of 10^7 molecules/ml (using antibody of the highest affinity found in practice), a non-competitive method is capable of yielding sensitivities some orders of magnitude greater than this. However, Fig. 4 also demonstrates that, assuming the use of high affinity antibodies (i.e. $\sim 10^{11}$ – 10^{12} l/M), maximal sensitivities yielded by isotopically based techniques (whether relying on labelled antibody (IRMA) or labelled analyte (RIA), or whether of competitive or non-competitive design) are closely comparable, i.e. of the order of 10^7 – 10^8 molecules/ml.

This limitation is a manifestation of the fact that, in the case of the non-competitive methods, an important constraint on assay sensitivity is (under certain circumstances) the 'specific activity' of the label used. On the other hand, limitation of assay sensitivity due to the low specific activity of radioisotopic labels does *not* often arise, in practice, in the case of competitive assays, whose sensitivity is generally restricted by other factors (Ekins, 1985). The fundamental significance of this conclusion is that, only by the use of labels possessing specific activities higher than those of the commonly used radioisotopes in assays of non-competitive design, can current

sensitivity limits be breached. Conversely, use of a higher specific activity label in a *competitive* assay will usually have no significant effect on its sensitivity (assuming experimental errors incurred in reagent manipulation of the magnitude generally encountered in practice).

High specific activity non-isotopic labels

The term 'specific activity' is conventionally applied, in the case of radioisotopic labels, to denote the number of radioactive disintegrations per unit time per unit weight of the isotope or labelled compound. In the present context, use of the term is widened to signify 'detectable events' per unit time per unit weight of labelled material. Thus it can be used to indicate the rate of photon emission by a chemiluminescent or fluorescent label, or the rate of conversion of substrate molecules—by an enzyme label—to molecules of a detectable product. The importance of the concept derives from the fact that 'signal measurement error' (i.e. error in the measurement of the label *per se*) is a contributory factor in limiting assay sensitivity, and may—when other sensitivity-constraining factors are reduced—become dominant. Furthermore, when extending the sensitivities of immunoassay systems beyond their present limits, the numbers of molecules involved are low, and statistical errors incurred in counting individual 'detectable events', and the time required to count them, may assume a particular importance.

Table 1 compares the specific activities of potentially useful labels with that of ^{125}I . All are of relevance in the context of this volume since chemiluminescent and fluorescent labels can be used to label antibodies (or antigens) directly; alternatively, enzyme labels catalysing reactions yielding chemiluminescent signals or fluorescent products can be utilized.

The importance of background in non-competitive immunoassays

A second important factor governing the sensitivity of non-competitive labelled-antibody immunoassays is the 'background' or 'blank' signal emitted in the absence of analyte, since error in the measurement of this signal is clearly a major determinant of the error in measurement of zero

Table 1. Relative specific activities of various isotopic and non-isotopic labels. Note that, though the specific activity of ^{125}I -labelled reagents does not, in practice, significantly limit the sensitivity of competitive assays (see Fig. 4), the lower specific activity of ^3H may severely restrict the sensitivity of competitive assays (e.g. of steroid hormones) which rely on the use of this particular radioisotope

Specific Activities

^{125}I :	1 detectable event/sec/ 7.5×10^6 labelled molecules.
^3H :	1 detectable event/sec/ 5.6×10^8 labelled molecules.
Enzymes:	Determined by enzyme 'amplification factor' and detectability of reaction product.
Chemiluminescent labels	1 detectable event/labelled molecule.
Fluorescent labels:	Many detectable events/labelled molecule.

dose. Amongst contributors to the background signal are the 'noise' of the measuring instrument itself, 'ambient' signal generators (such as, in 'sandwich' immunoassays, solid 'capture-antibody' supports or, in the case of radioisotopic methods, cosmic ray and other extraneous radiation sources) and 'non-specifically bound' labelled antibody. Minimization of each of these components is essential for maximal sensitivity: mere arithmetic subtraction of background is of absolutely no benefit in this context.

Non-specific binding of antibody is of particular interest, since the magnitude of this contribution is dependent, *inter alia*, on the amount of labelled antibody used in the system, and the duration of its exposure to analyte. Thus increasing the amount of labelled antibody increases the amount of such antibody bound to analyte; however, it may also increase the non-specifically bound moiety to a greater proportional extent, and thus cause a net reduction in sensitivity. This effect underlies the loss in sensitivity at higher antibody concentrations depicted in Fig. 5 (reproduced from Jackson *et al.*, 1983). This phenomenon also underlies the relationship between sensitivity and the affinity constant of the labelled antibody depicted in Fig. 4. The possession by labelled antibody of a high affinity constant implies that a

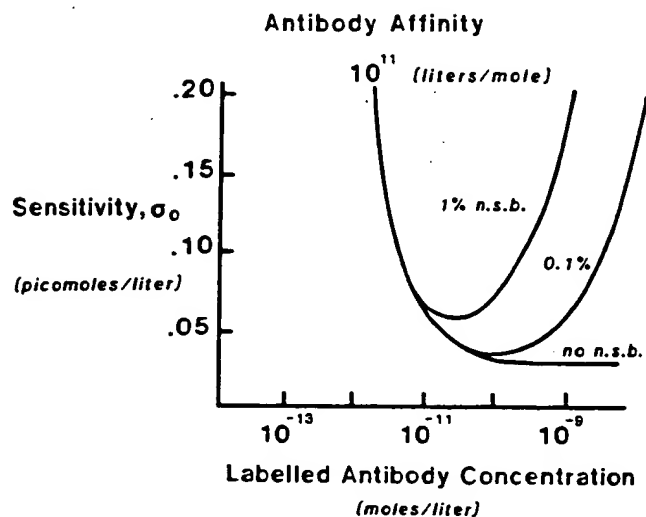


Figure 5. Assay sensitivity (represented by the standard deviation of the zero dose measurement, σ_0), plotted as a function of the concentration of labelled antibody (of affinity 10^{11} L/M) used in the assay, assuming different levels of non-specific binding of labelled antibody. (Note: an irreducible instrument background has been assumed in the computations represented; this limits the ultimate sensitivity attainable, regardless of the concentration of antibody used.)

lower concentration is required to yield the same level of analyte binding, albeit with reduced non-specific binding, thus increasing assay sensitivity

In summary, the high sensitivity of non-competitive labelled antibody methods derives essentially from their permitted use of optimal concentrations of antibody which (provided non-specific binding of labelled antibody is low) are generally considerably greater than in competitive methods, *not* from the fact that the antibody is labelled. Labelled antibody methods generally *fall* in sensitivity as the concentration of antibody is reduced towards zero, ultimately yielding a sensitivity theoretically identical to that of competitive methods (Rodbard and Weiss, 1973). (Paradoxically, early exponents of labelled antibody methods, whilst claiming them to be of higher sensitivity, also concluded that their sensitivity was *increased* by reduction in the amount of labelled antibody used (Woodhead *et al.*, 1971). This incorrect conclusion—based on observation of effects on the slope of the dose-response curve—exemplifies the many fallacies encountered in the immunoassay field stemming from confusion regarding the concept of sensitivity discussed above.) Finally it should be

emphasized that maximization of the sensitivity of a non-competitive immunoassay generally implies the selection of reagent concentrations and other experimental conditions such that the [analyte signal/background] ratio (i.e. s/b) is maximized. However, this simple relationship disregards statistical considerations which arise when the numbers of detectable events are very low, and a more appropriate objective may, under these circumstances, be maximization of the ratio s^2/b (Loevinger and Berman, 1951).

Other performance characteristics of competitive and non-competitive immunoassays

Non-competitive designs also display a number of other advantages deriving from the relatively high antibody concentrations on which they generally rely. These include increased reaction speeds (and hence shorter incubation times), decreased vulnerability to certain environmental effects (which cause variations in binding affinity between antibody and analyte), reduced sensitivity-dependence on high antibody binding affinity, etc.

Nevertheless a price has to be paid for these benefits; this includes the greater tendency of a large amount of antibody to bind molecules differing from, but with structural resemblance to, the analyte itself, implying a loss of assay *specificity*. This effect generally necessitates the use, whenever possible, of an 'immunoextraction' procedure using a second 'capture' antibody (usually directed against a different binding site, or 'epitope') as shown in Fig. 3(b). This technique—the 'sandwich' or 'two-site' immunoassay (Wide, 1971)—thus potentially combines the twin virtues of ultra-high sensitivity and specificity (together with short reaction time), features of crucial importance in many diagnostic situations (for example, in the detection of AIDS viral antigens). (Note, however, that the loss of specificity inherent in non-competitive assay designs implies that they are less readily applicable to the measurement of analytes of small molecular size, which cannot be simultaneously bound by two different antibodies directed against different antigenic sites on the molecule. Such analytes are generally more appropriately measured using 'competitive' assay methods.)

Development of ultra-sensitive immunoassay methodologies

The perception that the development of 'ultra-sensitive' immunoassay systems (i.e. systems surpassing conventional RIA methods in sensitivity) depends on (a) reliance on 'excess reagent' or 'non-competitive' assay designs; (b) the use of non-isotopic labels displaying higher specific activities than commonly used radioisotopes; (c) the development of efficient separation systems (ensuring minimization of non-specific antibody binding, and hence of signal 'backgrounds'), and (d) dual or multi-antibody analyte-recognition systems (exemplified by 'sandwich' or two-site assays) to maintain/increase assay specificity, has formed the basis of our own laboratory's immunoassay development since the early to mid-1970s (Ekins, 1978). This led us, *inter alia*, to an immediate recognition (Ekins, 1979, 1980) of the importance of the *in vitro* techniques of monoclonal antibody production pioneered by Köhler and Milstein (1975), which are currently the subject of bitter patent disputes in the USA (Ezzell, 1986, 1987a,b), and which may be expected in Europe.

Meanwhile, of the candidate labels for use in this context, both chemiluminescent and fluorescent labels offer many attractions. The development of stable, highly chemiluminescent, acridinium esters by McCapra and his colleagues (McCapra *et al.*, 1977) has subsequently been exploited by Weeks *et al.* (1983, 1984) and, more recently, by several commercial kit manufacturers; other workers have used more conventional chemiluminescent compounds to label immunoassay reagents (see, for example, Kohen *et al.*, 1984, 1985; Barnard *et al.*, 1985). Yet others have relied on enzyme labels to catalyse chemiluminescent (Whitehead *et al.*, 1983) and fluorogenic (Shalev *et al.*, 1980) reactions as indicated above. Detailed description of these various methodologies is presented by others in this volume and need not be duplicated here.

Common to all the 'ultra-sensitive' immunoassay methodologies relying on such alternative labels is their dependence on a non-competitive, labelled antibody, assay strategy whenever appropriate; however, for the reasons indicated above, *competitive* methods continue to be generally employed for the measurement of analytes of small molecular size (e.g. therapeutic drugs, steroid and thyroid hormones, etc.).

Nevertheless, the convenience (from a manufacturing viewpoint, and for other technical reasons) of relying on standard labelling procedures has meant that, even in these cases, labelled antibody techniques are increasingly preferred. Though the commercial kits based on these various labels differ to a minor extent in sensitivity, specificity, convenience, etc., such differences are at least partially attributable to differences in the physicochemical characteristics of the antibodies used in the kits, and to other 'immunological' factors unconnected with the particular nature of the label *per se*.

Despite the obvious attractions of chemiluminescent techniques in an immunoassay context, the use of fluorescent labels combined with sophisticated time-resolution techniques for their detection (a concept arising from discussions with J. F. Tait in 1970) appeared to us (in the mid-1970s) to offer more exciting long-term possibilities for a number of reasons. These naturally included attainment of the enhanced specific activities and high signal to background ratios required for ultra-sensitive immunoassay as indicated above. However, more importantly, fluorescence techniques also appeared to provide a simple route to the development of 'multi-analyte' assay systems of the kind described below.

In pursuance of this strategy, we began collaboration with LKB/Wallac, ca 1976-77, in the development of the instrumentation and technology required to develop such methods. Fortunately a group of fluorescent substances generally known as the lanthanide chelates (including, in particular, the chelates of europium, samarium and terbium facilitate such development, possessing prolonged fluorescence decay times (~ 10 - $1000 \mu\text{s}$), large Stokes shift ($\sim 300 \text{ nm}$) and other desirable physical characteristics which permit the construction of relatively cheap instrumentation for their measurement (Marshall *et al.*, 1981; Hemmilä *et al.*, 1983). The fluorescent properties of the lanthanide chelates may be compared with those of a conventional fluorophor such as fluorescein which is characterized by a much smaller Stokes shift ($\sim 28 \text{ nm}$), and a fluorescent decay time and emission spectrum which imply that it is less readily distinguished from fluorescent substances present in blood (such as bilirubin) or in plastic sample holders. The unique fluorescence characteristics of the lanthanide chelates thus permit them to be

measured in the presence of a fluorescence background (deriving from extraneous sources) which, in practice, approaches zero. Fig. 6 illustrates the basic concepts involved in pulsed-light, time-resolved, fluorescence measurement, which form the basis of the DELFIA immunoassay system currently marketed by LKB/Wallac.

Though it is inappropriate to pursue this subject in greater detail, attention should also be drawn to the possibilities offered by phase-resolved fluorimetry. This permits separate identification of fluorophores differing in fluorescence lifetime by their exposure to a sinusoidally modulated exciting light source, and observation of their demodulated, phase-shifted, light emission (McGown and Bright, 1984). This technique offers the possibility both of the development of homogeneous assays (relying on a difference in fluorescence decay time of bound and free forms of the fluorescent-labelled molecule), and of discriminating between two labelled antibodies in the context of multi-analyte 'ratiometric' immunoassay as discussed below.

'AMBIENT ANALYTE' IMMUNOASSAY

Before proceeding to a discussion of the development of multi-analyte assays, another important concept, termed 'ambient analyte immunoassay' (Ekins, 1983b), must first be examined. This term is intended to describe a type of immunoassay system which, unlike conventional

methods, measures the analyte *concentration* in the medium to which an antibody is exposed, being essentially independent both of sample volume, and of the amount of antibody present. This concept is illustrated in Fig. 7, and relies on the physicochemically-based proposition that, when a 'vanishingly small' amount of antibody (preferably, but not essentially, coupled to a solid support) is exposed to an analyte-containing medium, the resulting (fractional) occupancy of antibody binding sites solely reflects the ambient analyte concentration. Clearly the binding by antibody of analyte results in a depletion of the amount of analyte in the surrounding medium, but provided the proportion so bound is small (i.e. less than, for example, 1% of the total), such disturbance can be ignored. (This effect is closely analogous to that caused by the introduction of a thermometer into a medium possessing a much larger thermal capacity; the temperature disturbance caused by the thermometer itself is negligible and can, in these circumstances, be disregarded.)

The principles of ambient analyte assay derive from the recognition that *all* immunoassays essentially depend upon measurement of the 'fractional occupancy' by analyte of antibody binding sites following reaction of analyte with antibody as discussed above (Figs 3. (a) and (b)). The fractional occupancy of ('monospecific' or 'monoclonal') antibody binding sites in the presence of varying analyte concentrations, plot-

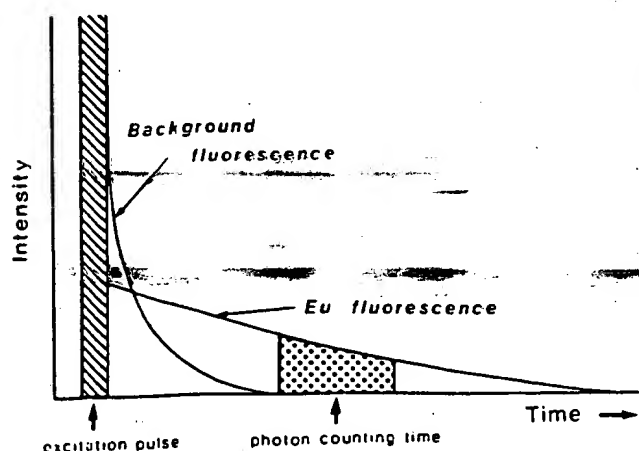


Figure 6. Basic principles of pulse-light, time resolved fluorescence. Fluorescence emitted by the fluorophore (typically a europium chelate) is distinguished from background fluorescence, which decays more rapidly.

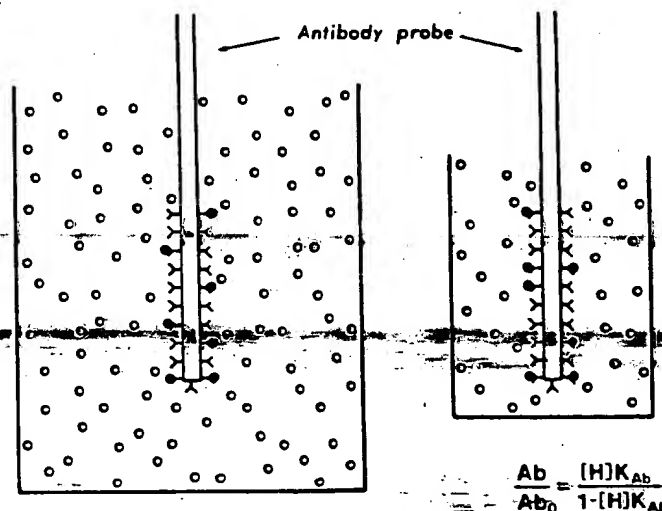


Figure 7. Basic principle of 'ambient analyte' immunoassay. The fractional occupancy (f) of a vanishingly small amount of antibody (of affinity K) is determined by the analyte concentration in the medium ($[A]$).

ted against antibody concentration, is portrayed in Fig. 8. The fraction of analyte bound is also plotted in this figure. (Note: for the sake of generality, all concentrations in this figure are expressed in terms of $1/K$, where K is the affinity constant of the antibody. For example, if $K = 10^{11}$ L/M, a concentration of $0.1 \times 1/K$ represents 0.1×10^{-11} M/L, or $0.1 \times 10^{-11} \times 10^{-3} \times 6.02 \times 10^{23} = 6.02 \times 10^8$ molecules/ml.)

It should be particularly noted that, at antibody concentrations of less than $ca 0.01 \times 1/K$ antibody fractional occupancy is essentially dependent solely on the analyte concentration in the medium, and is independent of variations in antibody concentration. This reflects the fact that this concentration of antibody binds less than approximately 1% of the analyte in the medium, irrespective of its concentration. This implies, for example, that the introduction of 10, 100, or 1000 antibody molecules into a medium containing billions of analyte molecules will result, in each case, in virtually identical fractional antibody binding-site occupancy, the upper limit of antibody concentration being determined by the antibody affinity constant. (An antibody concentration of $0.01 \times 1/K$ is a hundred-fold less than

that $(1 \times 1/K)$ necessary to bind 50% of a 'trace' amount of analyte (see Fig. 8), claimed by Berson and Yalow (1973) as maximizing assay 'sensitivity' (i.e. the slope of the dose-response curve when expressed in terms of bound/free labelled analyte). This false conclusion has subsequently become incorporated into the mythology of radioimmunoassay design which, regrettably, a majority of kit manufacturers continue to accept.)

The ambient analyte assay concept was originally exploited in the original development of what has come to be known as 'two-step' free hormone immunoassay (Ekins *et al.*, 1980), but it is clear that it is of far wider application, and can, in particular, be utilized in the construction of immunosensors and immunoprobes. One such example is a probe for the measurement of salivary steroids that is currently being developed in our laboratory. Comprising a small antibody-coated plastic 'dipstick' comparable in size and shape to a clinical thermometer, this device is intended to permit the measurement of salivary steroid levels without requiring the collection of saliva. However, the concept also underlies our approach to multi-analyte immunoassay, also under development in our laboratory.

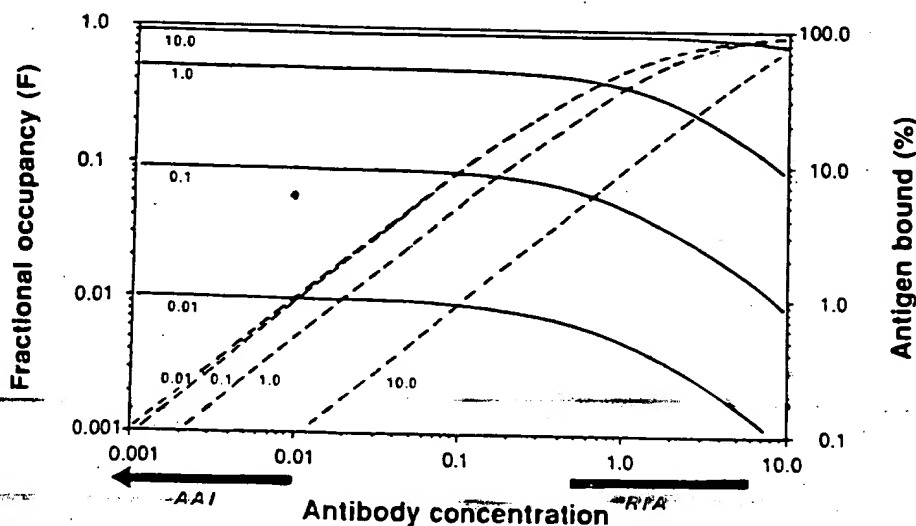


Figure 8. Fractional antibody binding-site occupancy (F) plotted as a function of antibody binding-site concentration for different values of analyte (antigen) concentration $[An]$. The percentage binding of analyte to antibody (b) is also shown. All concentrations are expressed in units of $1/K$. Note that for antibody concentrations of less than $0.01/K$ (approximately), concentration extending over several orders of magnitude, being governed solely by $[An]$. Note that radioimmunoassays and other 'competitive' immunoassays are commonly designed using antibody concentrations approximately $0.5/K$ – $1/K$ or above (implying $b_0 > 30\%$), in accordance with the precepts of Berson and Yalow (e.g. Berson and Yalow, 1973).

MULTI-ANALYTE 'RATIOMETRIC' IMMUNOASSAY SYSTEMS

The concepts relating to ambient analyte immunoassay and assay sensitivity outlined above are both exploited in our present development of a random access, multi-analyte, immunoassay technology capable of measuring, in the same small sample, virtually any number of individual analytes from selected analyte 'menus' (e.g. a hormone menu, viral antigen menu, an allergen menu, etc.). Many examples of a need to measure a multiplicity of different analytes in the same sample exist in medical diagnosis, for example, in the routine diagnosis of thyroid disease, where it is frequently necessary to measure a number of different hormones and thyroid-related proteins. At present, clinicians frequently experience difficulty in deciding on the best sequence of tests to arrive at a correct diagnosis. Such problems would be overcome were all relevant analytes measurable at a cost comparable to the cost of measurement of a single substance. Our own immediate objective is the development of a technology permitting the measurement of complete 'hormone profiles' using a single small blood sample. However, the need for 'multi-analyte', or 'random access' measurement is not confined to medical diagnosis: it also arises, for example, in the pharmaceutical industry (where there exists a requirement to ensure the purity of protein drugs synthesized by recombinant DNA techniques), in the food industry and elsewhere. Though still at an early stage, our approach to the achievement of this objective can be briefly indicated.

Multi-analyte assay: general principles

As discussed above, the notion of ambient analyte assay simultaneously introduces two extremely important and novel concepts: (a) that an estimate of analyte concentration can be based upon the use of an infinitesimal amount of 'sampling' antibody, and (b) that such an estimate derives from a direct measurement of fractional antibody occupancy by analyte, irrespective of the exact amount of antibody used. It should be emphasized that the latter proposition is valid only in the context of ambient analyte assay, and is *not* true in current conventional immunoassay systems (in which fractional antibody occupancy depends both upon the amount of antibody in the

system, and sample volume—see Fig. 8). In short, exposure of a small number of antibody molecules (in the form, for example, of a 'microspot' located on a solid support) to an analyte-containing fluid results in occupancy of antibody binding sites in the microspot reflecting the analyte concentration in the medium. Following such exposure, the antibody-bearing probe may be removed and exposed to a 'developing' solution containing a high concentration of an appropriate second antibody directed against either a second epitope on the analyte molecule if this is large (i.e. the occupied site), or against unoccupied antibody binding sites in the case of small analyte molecules (see Fig. 3(b)). (Note: an antibody simulating antigen, and reacting with unoccupied binding sites, is described as a 'mirror-image anti-idiotypic antibody'; the use of such an antibody instead of labelled antigen is convenient but not essential, and is suggested here merely to simplify illustration of the basic concepts involved.)

Subsequently, an estimate of binding-site occupancy of the 'sampling' (solid phase) antibody located in the microspot may be derived by measurement of the ratio of signals emitted by the two antibodies forming the dual-antibody 'couplets'. This can be conveniently achieved by labelling the 'sampling' and 'developing' antibodies with different labels, for example, a pair of radioactive, enzyme or chemiluminescent markers. Fluorescent labels are nevertheless particularly useful in this context because, by the use of optical scanning techniques, they permit arrays of different antibody 'microspots' distributed over a surface, each directed against a different analyte, to be individually examined, thus enabling multiple assays to be simultaneously carried out on the same small sample. Fig. 9 illustrates these basic ideas, and Fig. 10 such an array.

Microspot immunoassay sensitivity: theoretical considerations

The notion that it is, in principle, possible to measure an analyte concentration using a microspot of antibody comprising a number of antibody molecules in the range $ca\ 10^1$ – 10^6 is likely, at first sight, to appear surprising, and may, indeed, provoke scepticism regarding the assay sensitivities potentially attainable using this approach. Clearly a number of factors, such as the sensitivity

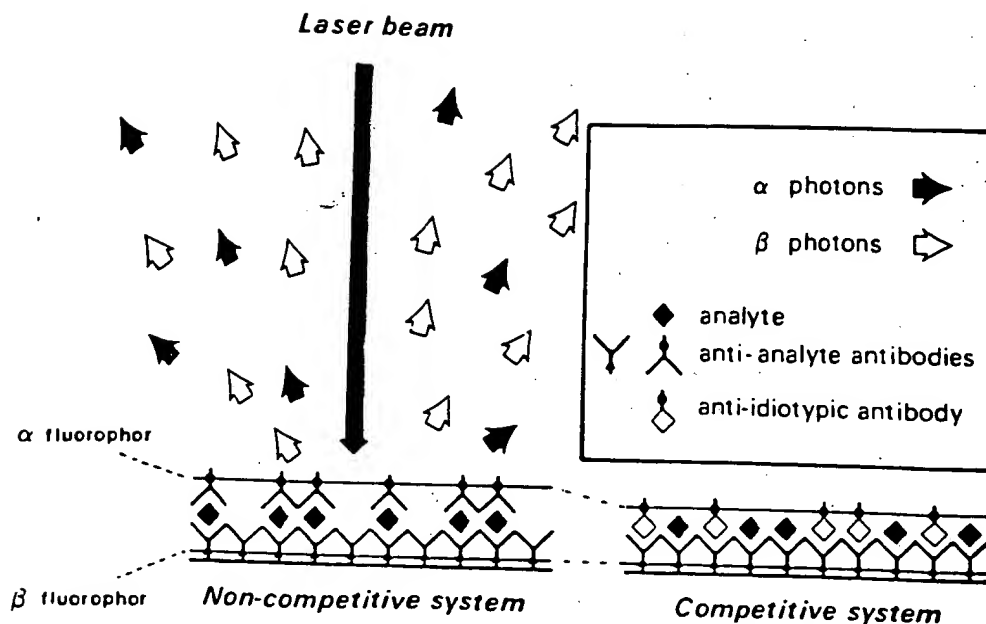


Figure 9. Basic principle of dual-label, ambient-analyte, immunoassay relying on fluorescent labelled antibodies. The ratio of α and β fluorescent photons emitted reflects the value of F (see Figs 5 and 6) and is solely dependent on the analyte concentration to which the probe has been exposed. It is unaffected by the amount or distribution of antibody coated (as a monomolecular layer) on the probe surface.

of the signal measuring equipment, the density of antibody molecules on the surface of the solid support, etc., are likely to play a part in determining final assay sensitivity. Such factors are, in turn, dependent on the efficiency with which the particular labels used can be detected, the adsorption properties of antibody supports,

etc. Though these are obviously variable, reasonable estimates can be made of the order of sensitivities likely to be achieved on the basis of some simple theoretical calculations. To clarify the following discussion, it is assumed that 'sensing' antibody can be uniformly and consistently coated on a solid matrix at a standard density, implying that only the 'developing' antibody need be labelled and measured in order to ascertain fractional occupancy of sensing antibody binding sites.

Fig. 11 illustrates the surface of an antibody microspot, of surface area $A(\mu\text{m}^2)$, and (uniformly) coated with antibody of affinity $K(\text{L/M})$ in a monomolecular layer of density $D(\text{molecules}/\mu\text{m}^2)$. Let us assume that the spot is exposed to an analyte-containing medium of volume $v(\text{ml})$, and containing an analyte concentration $C(\text{molecules}/\text{ml})$. The molecular concentration of antibody in the system is thus given by AD/v . (Note: the fact that antibody is situated on the surface of a solid support, and not evenly distributed throughout the medium, does not affect the extent of analyte binding at thermodynamic equilibrium, assuming that antibody binding sites are not impeded in their reactions and have not been damaged during the coating process.)

Meanwhile, fractional occupancy (F) of antibody-binding sites by analyte (at equilibrium) is

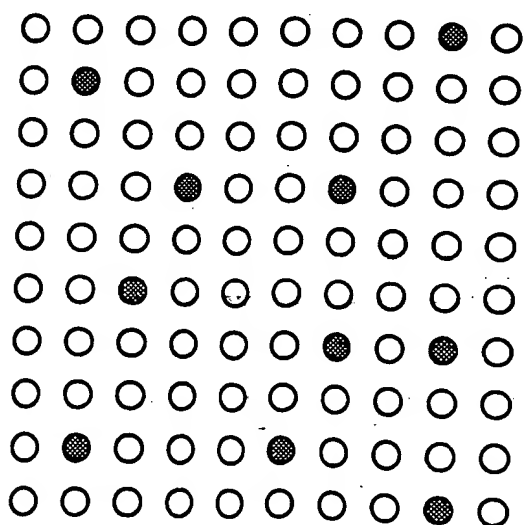


Figure 10. 'Multi-analyte' antibody array. Each antibody 'microspot' represents a 'vanishingly small' amount of antibody directed against an individual analyte.

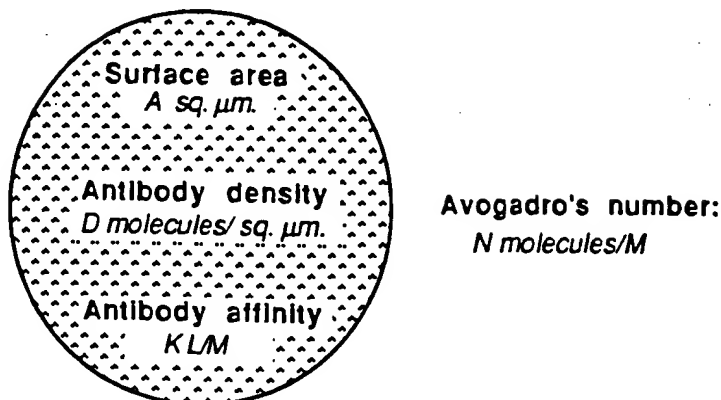


Figure 11. Microspot ambient-analyte immunoassay. The microspot shown is assumed to be uniformly coated with antibody, though if the dual-labelled antibody 'ratiometric' approach shown in Fig. 9 is adopted, uniform coating is not essential. The minimum fluid volume for ambient analyte assay conditions to prevail (enabling adoption of the ratiometric approach) is shown. Minimum test sample volume (M/S): $A \times D \times K \times 10^5/N$

given by the equation:

$$F^2 - F(1/q + p/q + 1) + p/q = 0 \quad (1)$$

where p = analyte concentration, q = antibody concentration (both expressed in units of $1/K$).

Thus, for antibody binding site concentrations $\rightarrow 0$ (i.e. $q < 0.01$), $F \approx p/(1 + p)$; (see Fig. 8).

Likewise, the fraction of analyte bound by antibody (f) at equilibrium is given by the equation:

$$f^2 - f(1/p + q/p + 1) + q/p = 0 \quad (2)$$

Thus, for analyte concentration $\rightarrow 0$ (i.e. $p < 0.01$), $f \approx q/(1 + q)$; (see Fig. 8). Furthermore, when $q < 0.01$, and when $p \geq 0$, $f < 0.01$.

Expressed in units of $1/K$; the concentration (q) in the assay of 'sensing' antibody situated on the microspot is given by $DAK/(\nu \times 6 \times 10^{20})$, (since Avogadro's constant, expressed as the number of molecules/mmol, is 6×10^{20} (approximately)). The fraction of an analyte concentration $\rightarrow 0$ which will be bound to the spot is therefore $DAK/(\nu \times 6 \times 10^{20} + DAK)$, implying that the number of analyte molecules bound to the spot is given by $\nu CDAK/(\nu \times 6 \times 10^{20} + DAK)$.

Case 1: sandwich (two-site) assay. Following incubation of sample with antibody, we assume the sample is removed, and the microspot then exposed to a volume V (ml) of a solution of a second, labelled, 'developing' antibody of affinity K^* (LM) at a concentration given by Q (expressed in units of $1/K^*$).

The fraction of analyte bound by labelled antibody (F^*) at equilibrium is given by the equation:

$$F^{*2} - F^*(1/P + Q/P + 1) + Q/P = 0 \quad (3)$$

where P represents the analyte concentration in the developing-antibody solution, expressed in units of $1/K^*$, i.e. $\nu CDAKK^*/[(\nu \times 6 \times 10^{20} + DAK)V \times 6 \times 10^{20}]$.

Assuming $P < 0.01$, $F^* \approx Q/(1 + Q)$. (For example, if $Q = 1$, the fraction of analyte molecules bound by labelled antibody = 0.5 approximately). Thus, since the number of analyte molecules bound to the spot is given by $\nu CDAK/(\nu \times 6 \times 10^{20} + DAK)$, the number of analyte molecules labelled by the second, developing, antibody is given by $\nu CDAKQ/[(\nu \times 6 \times 10^{20} + DAK)(1 + Q)]$, and the surface density of such molecules is given by $\nu CDKQ/[(\nu \times 6 \times 10^{20} + DAK)(1 + Q)]$. Moreover, assuming that $DAK \ll \nu \times 6 \times 10^{20}$ (i.e. that the amount of antibody in the system is such that 'ambient assay' conditions prevail, then the surface density (D^*) of developing-antibody molecules = $CDKQ/[(6 \times 10^{20})(1 + Q)]$ approximately. It should be noted that D^* is independent of both ν and V , also that the ratio $D^*/D = C \times KQ/[(6 \times 10^{20})(1 + Q)] = C \times \text{constant}$.

If the minimum detectable surface density of developing-antibody molecules (i.e. σ_{D^*} , the standard deviation of the measurement of D^* when $C = 0$) is given by D_{\min}^* (molecules/ μm^2) and C_{\min} represents the minimum detectable analyte concentration in the test sample, then,

disregarding non-specific binding of developing antibody within the microspot area,

$$C_{\min} = D_{\min}^* \times [(6 \times 10^{20})(1 + Q)]/DKQ \quad (4)$$

For example, if $Q = 1$, $D = 10^5$ molecules/ μm^2 , $K = 10^{11}$ L/M and $D_{\min}^* = 20$ molecules/ μm^2 , then $C_{\min} = 2.4 \times 10^6$ molecules/ml = 10^{-15} M/L. It should be noted, in this example, the fractional occupancy of the sensing antibody binding sites by the minimum detectable analyte concentration is 0.04%.

Case 2: anti-idiotypic antibody ('competitive') assay. In this case, we assume that, following removal of the sample, the microspot is exposed to a volume $V(\text{ml})$ of a solution of (for example) a second, labelled, anti-idiotypic antibody reacting with *unoccupied* sites on the sensing antibody. Using similar reasoning as above, we may likewise assume that the fraction of such sites which become occupied by the anti-idiotypic 'developing' antibody is given by $Q/(1 + Q)$, where Q is the developing-antibody concentration. However, the minimum detectable surface density of anti-idiotypic antibody is not, in a competitive design, the critical determinant of assay sensitivity; this parameter is essentially governed by the precision of the density measurement.

From Eq. (1), the fraction of sites *unoccupied* by analyte = $1/(1 + p)$, and the fraction occupied by anti-idiotypic antibody = $Q/(1 + p)(1 + Q)$. Thus, if the CV in the measurement of anti-idiotypic antibody is ϵ , the standard deviation is $\epsilon Q/(1 + p)(1 + Q)$. This term also represents the SD in the estimate of the fraction of sites *occupied* by analyte. Since the total number of antibody binding sites in the spot is DA , the SD in the estimate of occupied sites as $p \rightarrow 0$ (i.e. σD_0^*) approximates $\epsilon DAQ/(1 + Q)$; the SD in the occupied site surface-density estimate is thus $\epsilon DQ/(1 + Q)$. But the SD in the measurement of fractional binding-site occupancy when $p \rightarrow 0$ defines D_{\min} , and hence the minimum detectable analyte concentration in the test sample as indicated in Eq (4).

Thus

$$C_{\min} = D_{\min} \times [(6 \times 10^{20})(1 + Q)]/DKQ \quad (5)$$

$$= \epsilon DQ/(1 + Q) \pm [(6 \times 10^{20})(1 + Q)]/DKQ \quad (6)$$

$$= \epsilon/K \times (6 \times 10^{20}) \quad (7)$$

For example, if values of $Q = 1$, $D = 10^5$ molecules/ μm^2 , and $K = 10^{11}$ L/M are assumed as in the non-competitive example considered above, and the CV in the measurement of anti-idiotypic antibody density in the microspot is 1% (i.e. $\epsilon = 0.01$), then $D_{\min} = 500$ molecules/ μm^2 , and $C_{\min} = 6 \times 10^7$ molecules/ml = 10^{-13} M/L. Fractional occupancy of the sensing antibody binding sites by the minimum detectable analyte concentration is, in this example, 1%. It should be noted that the sensitivity limit of ϵ/K (expressed in molar terms) is identical to that previously established for conventional 'competitive' assays (Ekins and Newman, 1970), and which underlies the predictions represented in Fig. 4.

Such considerations appear to suggest (a) that microspot assay sensitivities superior to those obtainable by conventional radioisotopically based immunoassays are achievable, and (b) that sensitivities yielded by non-competitive microspot assays are likely to be considerably greater than those of corresponding competitive microspot assays. It must be emphasized, however, that, though such predictions are likely to prove correct, assumptions regarding the performance of the labels and signal-measuring instrument used are incorporated in the simple theoretical analysis discussed above. Such factors are clearly of importance in determining overall microspot immunoassay performance.

Practical implementation

The concepts discussed above are clearly exploitable using a variety of antibody labels, including chemiluminescent labels; however, our preliminary studies have been based on the use of conventional fluorophores, since the technology of simultaneous measurement of dual fluorescence from small areas is already well established. Because this volume centres on chemiluminescence, we shall provide only a brief indication of our initial experimental work in this area, which is currently based on the use of commercially available confocal microscopes.

Instrumentation: the laser scanning confocal microscope. In laser scanning confocal fluores-

ence microscopy, a small area of the specimen is illuminated by a focused laser beam; the fluorescence photons emanating solely from this area are, in turn, focused onto a photon detector. Both the intensity of illumination and the efficiency of light collection diminish rapidly with distance from the focal plane (Fig. 12). At the 'confocal' point, the projection of the illumination pinhole and the back-projection of the detector pinhole coincide. Such systems contrast with conventional epi-fluorescence methods, where the specimen is exposed to an essentially uniform flux of illumination (White *et al.*, 1987).

Sensitivity of current instruments. Typically, fluorescence photons emanating from the laser-

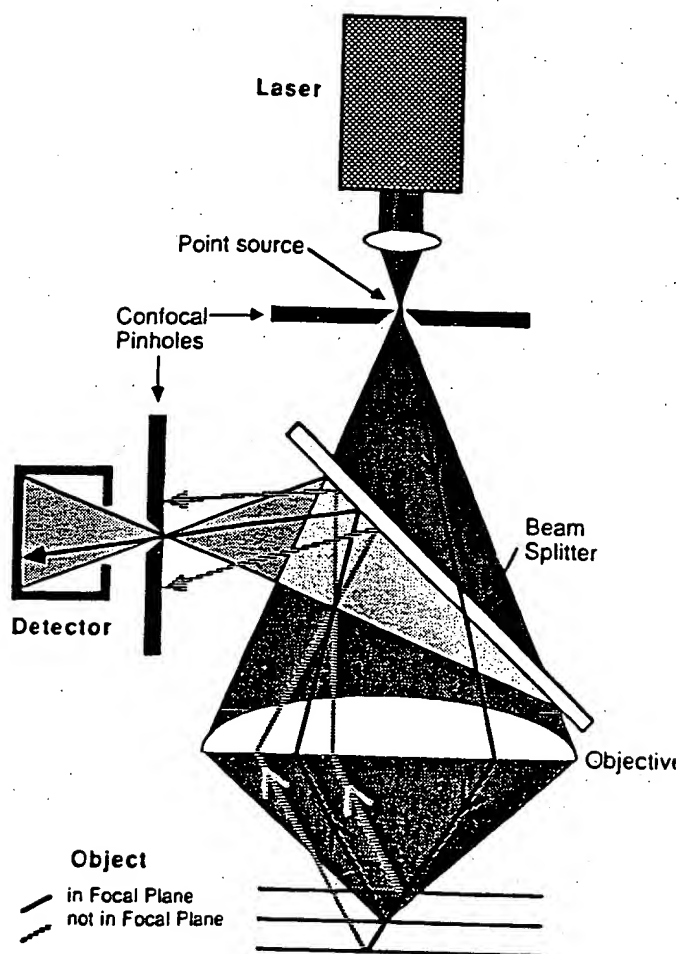


Figure 12. Principle of the confocal microscope. Illuminating light is focused at a point in the focal plane. Reflected light from this point is focused onto a detector. A complete two-dimensional image of structures within the focal plane is obtained by scanning the selected area of interest, and may be stored in a microcomputer for video display

illuminated area are detected by a low dark-current photomultiplier. Electrons spontaneously emitted by the photomultiplier photocathode contribute to the background signal of the instrument, and must, for highest sensitivity, be minimized. Fortunately the overall design of such instruments permits the photomultiplier photocathode to be of very small area, so that this particular source of background noise is not only small, but can be expected to reduce in relative importance with future improvement in photomultiplier design. Meanwhile current instruments already display very high sensitivity of detection of fluorescent signals. For example, the confocal microscope manufactured by Zeiss is claimed to display a lower detection limit for fluorescein of about ten molecules/ μm^2 (Ploem, 1986). Most commercially available FITC-labelled IgG attains a fluorophore/protein molar ratio of ~ 4 ; thus the detection limit (D_{\min}^*) of the Zeiss microscope is $\sim 2-3$ FITC-labelled IgG molecules/ μm^2 . This implies an analyte-concentration detection limit of $\sim 2.4 \times 10^5$ molecules/ml for a two-site assay, assuming the same parameter values as used in the examples discussed above, or 2.4×10^4 molecules/ml using a 'sensing' antibody of affinity 10^{12} L/M.

Another comparable instrument is the Bio-Rad/Lasersharp laser scanning confocal microscope, which we are currently using in the development of 'ratiometric' multi-analyte assay methodology in accordance with the principles outlined above (see Fig. 13). The argon laser in this system possesses two excitation lines at 488 and 514 nm. It is thus particularly efficient for the excitation of blue/green emitting fluorophores such as FITC (which displays an excitation maximum at 492 nm). However, it is considerably less efficient in the excitation of red-emitting fluorophores such as Texas red (excitation maximum 596 nm). However, the ratiometric immunoassay principle permits considerable variation in detection efficiencies of the two labels relied on since, *inter alia*, the specific activities of the two labelled antibody species forming the antibody couplets can be chosen to yield optimal signal ratios in the region of unity. Thus inefficiency of the argon laser in exciting red emitting fluorophores is not necessarily a major handicap in the present context.

Though the current Lasersharp instrument relies on a conventional microscope rather than a purpose-designed optical system (and appears to

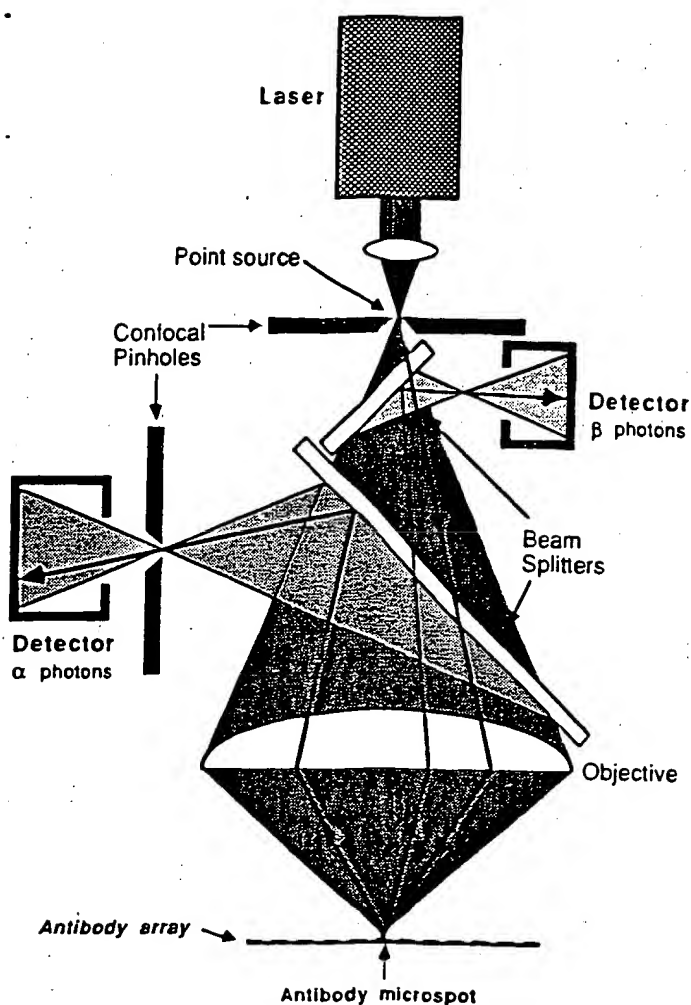


Figure 13. Dual-channel confocal fluorescence microscope permitting simultaneous measurement of the fluorescence signals from two fluorophors situated at the focal point. By scanning the antibody array, the ratio of signals from each antibody microspot may be determined

be less sensitive), it permits quantification of fluorescence signals generated from microspots of selected area. Initial studies have revealed that, under conditions that are not necessarily optimal, the instrument is capable of detecting approximately twenty-five FITC-labelled IgG molecules/ μm^2 , scanning an area of $\sim 50 \mu\text{m}^2$ (Fig. 14). It must be stressed that neither of these confocal microscopes are designed specifically for routine ratiometric multi-analyte immunoassay use, and it can be anticipated that future instruments constructed specifically for this purpose are likely to prove both cheaper and more sensitive.

Other instruments. The MPM 200 Microscope Photometer manufactured by Zeiss of West

Germany is anticipated to become available shortly. This photometer is claimed to be highly versatile: it can be used in transmission and reflection modes, and as a highly sensitive fluorimeter. The measuring field can be varied in shape and size for optimum adjustment to the specimen structure. More generally, the technology of sensitive light measurement is improving rapidly in response to needs in astronomy, the space program etc., such technology clearly being readily exploitable in a multi-analyte immunoassay context using light-generating labels in accordance with the broad principles presented here.

Solid antibody supports. On the basis of the theoretical considerations discussed above, it is evident that solid antibody supports for multi-analyte immunoassay use should display a capacity to adsorb a high surface density of antibody combined with low intrinsic signal-generating properties (for example, low intrinsic fluorescence), thus minimizing background. We have examined a number of materials, including polypropylene, Teflon, cellulose and nitrocellulose membranes and microtitre plates (clear polystyrene plates from Nunc; black, white and clear polystyrene plates from Dynatech with these criteria in mind. White Dynatech Microfluor microtitre plates, formulated specially for the detection of low fluorescence signals, yield high signal-to-noise ratios and have therefore been provisionally used in our developmental studies.

Surface density of antibody coating. Preliminary experiments using Microfluor plates have revealed that it is possible to coat them with antibody at a surface density of at least 5×10^4 IgG molecules/ μm^2 (Fig. 15). Moreover nearly all antibody molecules so deposited appear to retain immunological activity (Fig. 16).

Verification of the 'ratiometric' immunoassay concept. Our primary intention, in initial studies, has been establishment of the basic conditions which, using a particular instrument, can be anticipated on theoretical grounds to yield high assay sensitivity. Though the setting up of individual microspot immunoassays has thus appeared to us to be of secondary importance during the initial stages of our studies, we have nevertheless

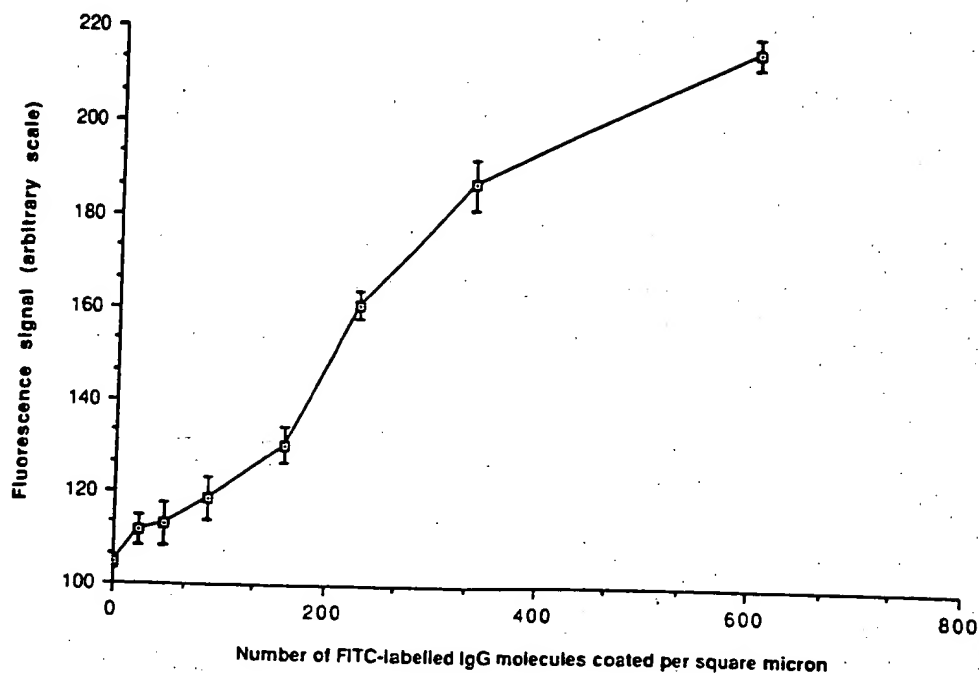


Figure 14. Fluorescence signal (arbitrary units), measured using the Bio-Rad/Laserssharp scanning confocal microscope, plotted as a function of the density of fluorescein-labelled IgG molecules (number of molecules/ μm^2) deposited on Dynatech Microfluor white microtitre plates

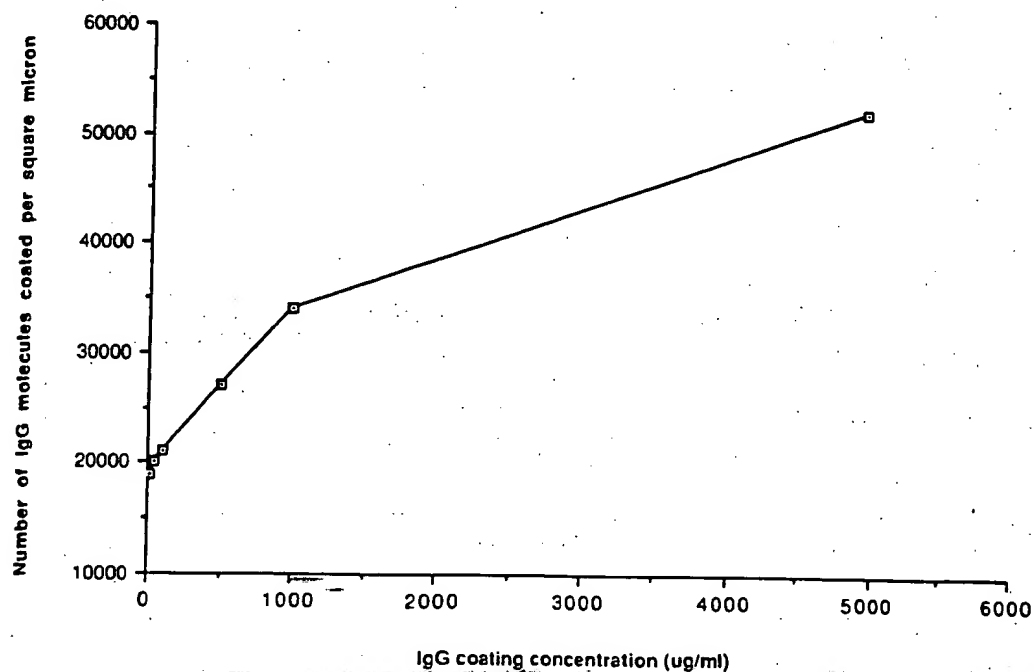


Figure 15. Surface density of IgG molecules (number of molecules/ μm^2) deposited on Dynatech Microfluor white plates plotted as a function of IgG concentration ($\mu\text{g/ml}$) in the coating solution

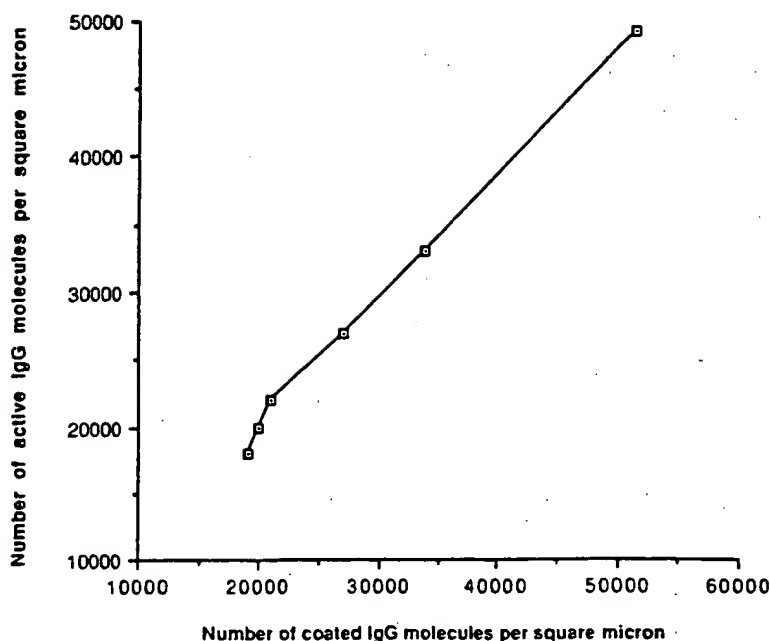


Figure 16. Surface density of immunoreactive IgG molecules (number of molecules/ μm^2) plotted as a function of the total surface density of IgG (number of molecules/ μm^2) on Dynatech Microfluor white microtitre plates

thought it useful to confirm the validity of our general concepts by comparing the performance of certain assays when constructed in microspot format and when conventionally designed. For example, we have compared a dual-labelled tumour necrosis factor (TNF) ratiometric assay system using Texas red and FITC-labelled antibodies with an optimized IRMA system using identical antibodies but with the second antibody ^{125}I -labelled. Although unoptimized, the ratiometric microspot assay yielded formal sensitivity values closely approaching that of the conventional, optimized, IRMA. Although verifying the general concepts underlying ratiometric microspot immunoassay methodology, further work is required to achieve the considerably greater sensitivity that theory predicts as achievable using optimized reagent concentrations and improved instrumentation.

CONCLUSION

As indicated above, differentiation of the fluorescent signals yielded by two fluorophores can be readily achieved solely on the basis of wavelength differences, and this approach has been relied on entirely in our preliminary studies. However,

other physical techniques exploiting differences in decay time of two or more fluorescence emissions (using, for example, a pulsed or sinusoidally modulated laser source, and time- or phase-resolving detectors) are available, and can be expected both to further reduce background and to improve signal resolution, thus increasing assay sensitivity and precision. These considerations aside, the basic technology involved closely resembles that employed in domestic compact disk recorders and other similar data-storage devices, the obvious difference being that light emitted from each of the discrete zones forming the antibody-array is fluorescent rather than reflected, and yields chemical rather than physical information. Indeed, our preliminary studies suggest that highly sensitive immunoassays using antibody microspots of surface area approximating $50\mu\text{m}^2$ are achievable, implying that some 2,000,000 different immunoassays could, in principle, be accommodated on a surface area of 1cm^2 . Though non-specific binding of a multiplicity of developing antibodies would probably prohibit the use of antibody arrays of this order, it is evident that the technology is capable of encompassing analyte numbers of the kind likely to be useful in practice.

The development of multi-analyte assay systems of this kind can be anticipated to bring about

fundamental changes in medical diagnosis and many other biologically related areas. Systems capable of measuring every hormone and other endocrinologically related substance within a single small sample of blood are within technological reach, providing data which, when analysed with the aid of computer-based 'expert' pattern-recognition systems, are likely to reveal endocrine deficiencies only dimly perceived using current 'single-analyte' diagnostic procedures. Such systems also provide a means to the development of a 'random access' immunoassay methodology, permitting the selection of any desired test or combination of tests from an extensive analyte menu. Clearly the accommodation of a wide range of individual immunoassays on a small immunoprobe (comparable in its overall physical dimensions with a few drops of blood) is likely to totally transform the logistics of immunodiagnostic testing, and genuinely represents, in our view, 'next generation' immunoassay methodology.

Acknowledgement

These studies are being generously supported by The Wolfson Foundation.

REFERENCES

- Barnard, G. J. R., Kim, J. B. and Williams, J. L. (1985). Chemiluminescence immunoassays and immunochemiluminometric assays. In *Alternative Immunoassays*, Collins, W. P. (Ed.), John Wiley, Chichester, pp. 123-152.
- Berson, S. A. and Yalow, R. S. (1973). Measurement of hormones—radioimmunoassay. In *Methods in Investigative and Diagnostic Endocrinology*, 2A, Berson, S. A. and Yalow, R. S. (Eds), North Nolland/Esevier, New York, pp. 84-135.
- Dakubu, S., Ekins, R., Jackson, T. and Marshall, N. J. (1984). High sensitivity, pulsed light time-resolved fluoroimmunoassay. In *Practical Immunoassay. The State of the Art*, Butt, W. R. (Ed.), Marcel Dekker, New York, pp. 71-101.
- Ekins, R. P. (1976). General principles of hormone assay. In *Hormone Assays and their Clinical Application*, Loraine, J. A. and Bell, E. T. (Eds), Churchill Livingstone, Edinburgh, pp. 1-72.
- Ekins, R. P. (1978). The future development of immunoassay. In *Radioimmunoassay and Related Procedures in Medicine 1977*, IAEA, Vienna, pp. 241-275.
- Ekins, R. P. (1979). Radioassay methods. In *Radiopharmaceuticals II: Proceedings, 2nd International Symposium on Radiopharmaceuticals*, 19-22 March 1979, Seattle, Washington, Sorenson, J. A. (Ed), Society of Nuclear Medicine, New York, pp. 219-240.
- Ekins, R. P. (1980). More sensitive immunoassays. *Nature*, **284**, 14-15.
- Ekins, R. P. (1983a). The precision profile: its use in assay design, assessment and quality control. In *Immunoassays for Clinical Chemistry*, Hunter, W. M. and Corrie, J. E. T. (Eds), Churchill Livingstone, Edinburgh, pp. 76-105.
- Ekins, R. P. (1983b). Measurement of analyte concentration. *British Patent* no. 8224600.
- Ekins, R. (1985). Current concepts and future developments. In *Alternative Immunoassays*, Collins, W. P. (Ed.), John Wiley, Chichester, pp. 219-237.
- Ekins, R. P. and Newman, B. (1970). Theoretical aspects of saturation analysis. In *Karolinska Symposia on Research Methods in Reproductive Endocrinology. 2nd Symposium: Steroid Assay by Protein Binding*, Diczfalussy, E. (Ed.), The Reproductive Endocrinology Research Unit, Karolinska sjukhuset Stockholm, pp. 11-36.
- Ekins, R. P., Newman, B. and O'Riordan, J. L. H. (1968). Theoretical aspects of 'saturation' and radioimmunoassay. In *Radioisotopes in Medicine: In Vitro Studies*, Hayes, R. L., Goswitz, F. A. and Murphy, B. E. P. (Eds), Oak Ridge Symposia, USAEC, Oak Ridge, Tennessee, pp. 59-100.
- Ekins, R. P., Newman, B. and O'Riordan, J. L. H. (1970a). Saturation assays. In *Statistics in Endocrinology*, McArthur, J. W. and Colton, T. (Eds), MIT Press, Cambridge, MA, pp. 345-378.
- Ekins, R. P., Newman, B. and O'Riordan, J. L. H. (1970b). Competitive protein-binding assays. Discussion. In *Statistics in Endocrinology*, McArthur, J. W. and Colton, T. (Eds), MIT Press, Cambridge, MA, pp. 379-392.
- Ekins, R. P., Filetti, S., Kurtz, A. B. and Dwyer, K. (1980). A simple general method for the assay of free hormones (and drugs); its application to the measurement of serum free thyroxine levels and the bearing of assay results on the 'free thyroxine' concept. *J. Endocrinol.*, **85**, 29-30.
- Ezzell, C. (1986). Hybritech versus Abbott. *Nature*, **324**, 506.
- Ezzell, C. (1987a). Judge confirms injunction in sandwich assay patent suit. *Nature*, **326**, 532.
- Ezzell, C. (1987b). Hybritech wins court injunction over sandwich assays. *Nature*, **327**, 5.
- Hemmilä, I., Dakubu, S., Mikkala, V.-M., Siiteri, H. and Lovgren, T. (1983). Europium as a label in time-resolved immunofluorometric assays. *Anal. Biochem.*, **137**, 335-343.
- Jackson, T. M., Marshall, N. J. and Ekins, R. P. (1983). Optimisation of immunoradiometric (labelled antibody) assays. In *Immunoassays for Clinical Chemistry*, Hunter, W. M. and Corrie, J. E. T. (Eds), Churchill Livingstone, Edinburgh, pp. 557-575.
- Kohen, F., Bayer, E. A., Wilchek, M., Barnard, G., Kim, J. B., Collins, W. P., Beheshti, I., Richardson, A. and McCapra, F. (1984). Development of luminescence-based immunoassays for haptens and for peptide hormones. In *Analytical Applications of Bioluminescence and Chemiluminescence*, Kricka, L., Stanley, P. E., Thorpe, G. H. G. and Whitehead, T. P. (Eds), Academic Press, New York, pp. 149-158.
- Kohen, F., Pazzagli, M., Serio, M., DeBoever, J. and Vanderkerckhove, D. (1985). Chemiluminescence and bioluminescence immunoassay. In *Alternative Immunoassays*, Collins, W. P. (Ed), John Wiley, Chichester, pp. 103-121.
- Köhler, G. and Milstein, C. (1975). Continuous culture of

- fused cells secreting specific antibody. *Nature*, 256, 495-497.
- Loevinger, R. and Berman, M. (1951). Efficiency criteria in radioactive counting. *Nucleonics*, 9, 26.
- Marshall, N. J., Dakubu, S., Jackson, T. and Ekins, R. P. (1981). Pulsed-light, time-resolved, fluoroimmunoassay. In *Monoclonal Antibodies and Developments in Immunoassay*, Albertini, A. and Ekins, R. (Eds), Elsevier/North Holland, Amsterdam, pp. 101-108.
- McCapra, F., Tutt, D. E. and Topping, R. M. (1977). Assay method utilizing chemiluminescence. *British Patent no.* 1, 461, 877.
- McGown, L. B. and Bright, F. V. (1984). Phase-resolved fluorescence spectroscopy. *Anal. Chem.*, 56, 1400-1417.
- Miles, L. E. H. and Hales, C. N. (1968). An immunoradiometric assay of insulin. In *Protein and Polypeptide Hormones, Pt. 1*, Margoulies, M. (Ed.), Excerpta Medica, Amsterdam, pp. 61-70.
- Ploem, J. S. (1986). New instrumentation for sensitive image analysis of fluorescence in cells and tissues. In *Applications of Fluorescence in the Biological Sciences*, Tayer, D. L., Waggoner, A. S., Lanni, F., Murphy, R. and Birge, R. (Eds), Alan R. Liss, New York, pp. 289-300.
- Rodbard, D. and Weiss, G. H. (1973). Mathematical theory of immunometric (labelled antibody) assay. *Analyt. Biochem.*, 52, 10-44.
- Shalev, A., Greenberg, G. H. and McAlpine, P. J. (1980). Detection of attograms of antigen by a high sensitivity enzyme-linked immunosorbent assay (HS-ELISA) using a fluorogenic substrate. *J. Immunol. Methods*, 38, 125-139.
- Tait, J. F. (1970). Competitive protein-binding assays. Discussion. In *Statistics in Endocrinology*, McArthur, J. W. and Colton, T. (Eds), MIT Press, Cambridge, MA. pp. 379-392.
- Weeks, I., McCapra, F., Campbell, A. K. and Woodhead, J. S. (1983). Immunoassays using chemiluminescent labelled antibodies. In *Immunoassays for Clinical Chemistry*, Hunter, W. M. and Corrie, J. E. T. (Eds), Churchill Livingstone, Edinburgh, pp. 525-530.
- Weeks, I., Campbell, A. K., Woodhead, S. and McCapra, F. (1984). Immunoassays using chemiluminescent labels. In *Practical Immunoassay. The State of the Art*, Butt, W. R. (Ed.), Marcel Dekker, New York, pp. 103-116.
- White, J. G., Amos, W. B. and Fordham, M. (1987). An evaluation of confocal versus conventional imaging of biological structures by fluorescence light microscopy. *J. Cell Biol.*, 105, 41-48.
- Whitehead, T. P., Thorpe, G. H., Carter, T. J., Groucutt, C. and Kricka, L. J. (1983). Enhanced luminescence procedure for sensitive determination of peroxidase-labelled conjugates in immunoassay. *Nature*, 305, 158-159.
- Wide, L. (1971). Solid phase antigen-antibody systems. In *Radioimmunoassay Methods*, Kirkham, K. E. and Hunter, W. M. (Eds), Churchill Livingstone, Edinburgh, pp. 405-418.
- Woodhead, J. S., Addison, G. M., Hales, C. N. and O'Riordan, J. L. H. (1971). Discussion. In *Radioimmunoassay Methods*, Kirkham, K. E. and Hunter, W. M. (Eds), Churchill Livingstone, Edinburgh, pp. 467-488.
- Yalow, R. S. and Berson, S. A. (1970a). Radioimmunoassays. In *Statistics in Endocrinology*, McArthur, J. W. and Colton, T. (Eds), MIT Press, Cambridge, MA. pp. 327-344.
- Yalow, R. S. and Berson, S. A. (1970b). Competitive protein-binding assays. Discussion. In *Statistics in Endocrinology*, McArthur, J. W. and Colton, T. (Eds), MIT Press, Cambridge, MA. pp. 379-392.

JOURNAL OF
BIOLUMINESCENCE AND CHEMILUMINESCENCE

**Bioluminescence and Chemiluminescence:
Studies and Applications in
Biology and Medicine**

Proceedings of the Vth International
Symposium on Bioluminescence and
Chemiluminescence

Editors:

M. Pazzagli, E. Cadenas, L. J. Kricka,
A. Roda and P. E. Stanley

Volume 4 1989

 **WILEY**

Chichester · New York · Brisbane · Toronto · Singapore

JBCHE7 4(1) 1-646
ISSN 0884-3996

CLIN. CHEM. 37/11, 1955-1967 (1991)

BY COPYRIGHT LAW

Multianalyte Microspot Immunoassay—Microanalytical "Compact Disk" of the Future

R. P. Ekins and F. W. Chu

Throughout the 1970s, controversy centered both on immunoassay "sensitivity" per se and on the relative sensitivities of labeled antibody (Ab) and labeled analyte methods. Our theoretical studies revealed that RIA sensitivities could be surpassed only by the use of very high-specific-activity nonisotopic labels in "noncompetitive" designs, preferably with monoclonal antibodies. The time-resolved fluorescence methodology known as DELFIA—developed in collaboration with LKB/Wallac—represented the first commercial "ultrasensitive" nonisotopic technique based on these theoretical insights, the same concepts being subsequently adopted in comparable methodologies relying on the use of chemiluminescent and enzyme labels. However, high-specific-activity labels also permit the development of "multianalyte" immunoassay systems combining ultrasensitivity with the simultaneous measurement of tens, hundreds, or thousands of analytes in a small biological sample. This possibility relies on simple, albeit hitherto-unexploited, physicochemical concepts. The first is that all immunoassays rely on the measurement of Ab occupancy by analyte. The second is that, provided the Ab concentration used is "vanishingly small," fractional Ab occupancy is independent of both Ab concentration and sample volume. This leads to the notion of "ratiometric" immunoassay, involving measurement of the ratio of signals (e.g., fluorescent signals) emitted by two labeled Abs, the first (a "sensor" Ab) deposited as a microspot on a solid support, the second (a "developing" Ab) directed against either occupied or unoccupied binding sites of the sensor Ab. Our preliminary studies of this approach have relied on a dual-channel scanning-laser confocal microscope, permitting microspots of area $100\ \mu\text{m}^2$ or less to be analyzed, and implying that an array of 10^6 Ab-containing microspots, each directed against a different analyte, could, in principle, be accommodated on an area of $1\ \text{cm}^2$. Although measurement of such analyte numbers is unlikely ever to be required, the ability to analyze biological fluids for a wide spectrum of analytes is likely to transform immunodiagnostics in the next decade.

Additional Keyphrases: *ratiometric immunoassays* • *scanning-laser confocal microscope* • *fluoroimmunoassay*

Immunoassay and other protein-binding assay methods based on the use of radioisotopic labels have played a major role in medicine during the past three decades.

Their utility and importance have derived primarily from the structural specificity of many reactions between binding proteins and analytes and the detectability of isotopically labeled reagents, the latter endowing such techniques with "exquisite sensitivity." Recently, however, interest has increasingly focused on nonisotopic techniques based on identical analytical principles, differing only in the nature of the marker used to label the reactant (e.g., antibody or antigen), whose distribution between reacted ("bound") and unreacted ("free") fractions constitutes the assay "response."

The basic aims underlying this interest can be broadly classed under four main headings:

- avoidance of the environmental, legal, economic, and practical disadvantages of isotopic techniques (e.g., limited shelf life of isotopically labeled reagents, problems of radioactive waste disposal, cost and complexity of radioisotope counting equipment), particularly those impeding the development of, for example, simple diagnostic kits for home or doctor's office use;
- achievement of greater assay sensitivity;
- "direct" measurement of analyte concentrations by use of transducer-based "immunosensors";
- simultaneous measurement of multiple analytes ("multianalyte assay").

In this presentation I will focus primarily on the last of these objectives, using this to set out the principles underlying our present attempts to develop a new "miniaturized" technology that will permit the simultaneous measurement of an unlimited number of analytes in a small biological sample such as a single drop of blood. However, retention (and, if possible, improvement) of the high sensitivities of conventional isotopic techniques is a basic aim not only of our own studies in this area but also of most other endeavors falling under the above headings. It is therefore appropriate to preface this paper with a discussion of the general principles underlying the attainment of high binding-assay sensitivity.

Immunoassay Sensitivity: Some Basic Concepts

Definition of Assay Sensitivity

The need to establish assay conditions yielding maximal sensitivity underlay the independent construction of mathematical theories of immunoassay design by both Yalow and Berson (1) and Ekins et al. (2) in the course of the original development of these methods in the early 1960s. Regrettably, these theoretical studies led to a prolonged controversy, arising largely from the conflicting concepts of "sensitivity" adopted by the two groups (see Figure 1). Briefly, Berson and Yalow, in their many publications relating to immunoassay design (e.g., 1, 3), defined sensitivity as the slope of the

Department of Molecular Endocrinology, University College and Middlesex School of Medicine, Mortimer St., London W1N 8AA, U.K.

Presented at the 23rd annual Oak Ridge Conference on Advanced Analytical Concepts for the Clinical Laboratory, St. Louis, MO, April 1991.

Received May 8, 1991; accepted August 20, 1991.

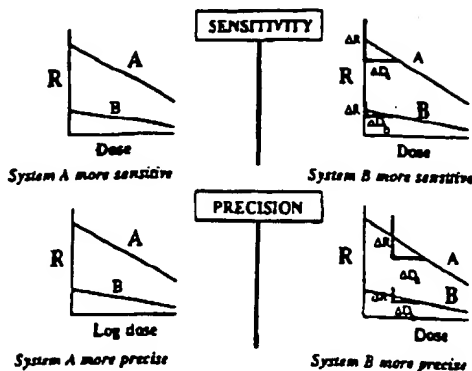


Fig. 1. The differing concepts of sensitivity and precision underlying radioimmunoassay design theories developed by (left) Yalow and Berson (e.g., 1, 3) and (right) Ekins et al. (2, 4)

Yalow and Berson define assay A as more sensitive because it yields a response curve of greater slope. Ekins et al. define assay B as more sensitive because the imprecision of measurement of zero dose (σ_0) is less. Yalow and Berson likewise define an assay system as more precise if it yields a steeper response curve when data are plotted on a log dose scale

response curve relating the fraction or percentage of labeled antigen bound (b) to analyte concentration ($[H]$). In contrast, Ekins et al. (e.g., 2, 4) defined sensitivity as the (im)precision of measurement of zero dose, this quantity being indicative of, and essentially equivalent to, the lower limit of detection.

The key difference between these two definitions clearly lies in the dependence of the assay detection limit on the error (imprecision) in the measurement of the response variable. By neglecting this crucial factor, the "response curve slope" definition leads to many obvious absurdities. For example, plotting conventional RIA data in terms of the response metameter B/F (i.e., the bound to free ratio) suggests that assay "sensitivity" is increased by increasing the antibody concentration in the system; however, the converse conclusion is reached if identical data are plotted in terms of F/B (see Figure 2). Observation of the shape and slopes of response curves without detailed error analysis thus constitutes a totally misleading guide to optimal immunoassay design. This approach has, however, characterized many of the studies conducted in the immunoassay field during the past 30 years, and has been the source of much

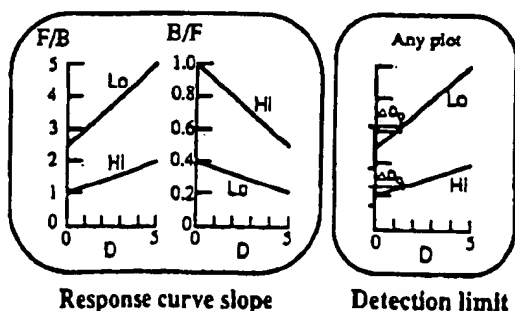


Fig. 2. Schematic representation of RIA dose-response curves observed for high and low antibody concentrations plotted in terms of (left) the free/bound fraction (F/B); (center) the bound/free fraction (B/F)

Note that the low antibody concentration yields a response curve of greater slope when the assay response is plotted in terms of F/B , but of lower slope when plotted in terms of B/F . The precision of measurement of zero dose (ΔD_0) is independent of the coordinate frame used to plot assay data (see right)

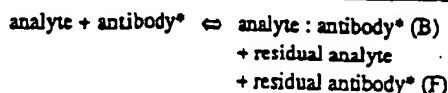
mythology. For example, consideration of the Law of Mass Action reveals that, when response curves corresponding to different antibody concentrations are plotted in terms of b vs $[H]$, the maximal slope at zero dose is obtained for a concentration of $0.5/K$ (where K is the affinity constant), in which circumstance the zero dose response (b_0) is 33%. This conclusion led to Berson and Yalow's enunciation of the well-known dictum (which, albeit erroneous, is broadly adhered to by many immunoassay practitioners and kit manufacturers) that, to maximize RIA sensitivity, the amount of antibody to use in the system is that which binds 33% of labeled antigen in the absence of unlabeled antigen (1, 3).

Disagreement regarding the concept of sensitivity inevitably led to prolonged dispute regarding immunoassay design (5). However, although it is still common to encounter publications in the field that rely solely on the response curve slope as a measure of sensitivity, the assay detection limit is now widely accepted as the only valid indicator of this parameter, and we do not therefore intend to dwell further on this issue here. It is nevertheless relevant to an understanding of the "miniaturized" assay methodology described below to emphasize that untenable concepts of both sensitivity and precision underlie many of the commonly accepted rules governing current immunoassay-design practice, some of which are contravened in our own approach.

Basic Immunoassay Designs

It is likewise important in the present context to comprehend the basis of the various types of immunoassays currently in use, and the constraints on the sensitivities of which they are potentially capable. The radioimmunoassay and analogous protein-binding assay techniques originally developed for the measurement of insulin by Yalow and Berson (6), and of thyroxine and vitamin B_{12} by Ekins and Barakat (7, 8), relied on the use of a labeled analyte marker to reveal the products of the binding reactions between analyte and binder (Figure 3, left). This approach has subsequently often been portrayed as relying on "competition" between labeled and unlabeled analyte molecules for a limited number of protein-binding sites, such assays being frequently referred to as "competitive."

Subsequently, Wide et al. in Sweden (9), followed shortly by Miles and Hales in the U.K. (10), developed labeled antibody methods (Figure 3, right). These methods represented an extension of the "labeled reagent" methods (utilizing radiolabeled organic compounds such as ^{131}I -labeled p -iodosulfonyl chloride, ^3H acetic anhydride, and other similar reagents) devised, during the early 1950s, by Keston et al. (11), Avivi et al. (12), and others for quantifying amino acids, steroid and thyroid hormones, etc. Although radiolabeled antibody methods (immunoradiometric assays; IRMAs) were originally claimed (13) to be more sensitive than methods based on the use of radiolabeled analyte, these claims were supported by neither rigorous theoretical analysis nor persuasive experimental evidence, and for some time remained controversial. Further doubt on their validity



Measure "fraction bound" (B)	$[\text{Ab}]_{\text{opt}} \rightarrow \infty$
Measure "fraction free" (F)	$[\text{Ab}]_{\text{opt}} \rightarrow 0$

Fig. 3. Labeled-analyte (left) and labeled-antibody (right) assay systems compared

Labeled-analyte assay systems essentially rely on observation of an analyte "marker" to reveal the products of the reaction between analyte and antibody (although the labeled analyte is not necessarily identical to the unlabeled analyte in its binding characteristics vis-à-vis antibody). Note that, irrespective of which fraction of the labeled analyte is measured after the binding reaction, the optimal antibody concentration required to maximize sensitivity in such a system tends toward zero (assuming a background signal of 0). Labeled-antibody systems rely on observation of an antibody "marker" to reveal the products of the binding reaction between analyte and antibody. In this case, the optimal antibody concentration required to maximize sensitivity tends toward zero when the "free" antibody fraction is measured, but tends toward infinity when the bound fraction is determined (likewise assuming zero background).

was cast by the publication by Rodbard and Weiss in 1973 (14) of detailed theoretical studies demonstrating that both labeled analyte and labeled antibody methods possessed essentially equal sensitivities. (Note: These authors suggested that IRMAs might be more sensitive in the assay of small polypeptides, in which radioiodine incorporation into the antigen molecule was restricted; conversely, these assays would be less sensitive for the measurement of antigens of high molecular mass.) Nevertheless, despite the appearance of this publication, the belief that labeled antibody methods per se are intrinsically more sensitive than the corresponding labeled analyte methods gained wide acceptance among clinical chemists.

The reason for confusion on this issue is that the greater potential sensitivity of certain assay formats is not really a consequence of the labeling of antibody as opposed to analyte; indeed, the apparent antithesis between labeled-analyte and labeled-antibody methods diverts attention from the true reasons underlying the superior sensitivity of certain assay designs. Theoretical analysis (see, e.g., 4, 15) reveals that, assuming "perfect" separation of the products of the binding reaction (i.e., no misclassification of bound and free moieties), the optimal antibody concentration (for maximal sensitivity) in a labeled analyte immunoassay invariably tends to zero, irrespective of whether the free or bound labeled analyte fraction is measured, whereas in labeled-antibody methods the optimal antibody concentration depends on which labeled-antibody fraction is measured (see Figure 3). If the free (unreacted) antibody fraction is measured, the optimal concentration also tends to zero; conversely, if the analyte-bound fraction is measured, the concentration tends to infinity. In short, of the four basic measurement strategies available—labeled analyte, with measurement of free or bound reaction product, and labeled antibody, also with measurement of free or bound product—only one permits, in practice, the use of antibody concentrations approaching infinity.

This particular approach may, for want of a better term, be described as "noncompetitive," although it must be emphasized that such terminology involves a departure from the original meanings attached to "competitive" and "noncompetitive" when these descriptions were first used in the present context. Indeed, as discussed below, assays may be subclassified in this manner when no labeled reagent of any kind is involved.

However, the categorization of immunoassays and other binding assays as competitive or noncompetitive, depending on the binding agent concentration yielding maximal assay sensitivity, itself obscures the underlying reasons for the existence of this divergence in assay designs, and may thus be misleading. These reasons may be more readily understood if the basic principles of such assays are portrayed differently from their customary presentation.

The "Antibody Occupancy Principle" of Immunoassay

When a "sensor" antibody is introduced into an analyte-containing medium, binding sites on the antibody are occupied by analyte molecules to a fractional extent that reflects both the equilibrium constant governing the binding reaction, and the final concentration of free analyte present in the mixture. This proposition stems immediately from the Law of Mass Action, which can be written as

$$[\text{AbAg}]/[\text{Ab}] = K[\text{fAg}] \quad (1)$$

or as fractional occupancy of antibody binding sites, given by

$$[\text{AbAg}]/[\text{Ab}] = K[\text{fAg}]/(1 + K[\text{fAg}]) \quad (2)$$

where $[\text{AbAg}]$, $[\text{Ab}]$, $[\text{fAb}]$, and $[\text{fAg}]$ represent the concentrations (at equilibrium) of bound and total antibody, and free antibody and antigen (analyte), respectively, and K = equilibrium constant. The final concentration of free analyte generally depends on the concentrations of both total analyte and antibody; however, when total antibody approximates $0.05/K$ or less, free and total antigen ($[\text{Ag}]$) concentrations do not differ significantly, and fractional occupancy of antibody is given by

$$[\text{AbAg}]/[\text{Ab}] = K[\text{Ag}]/(1 + K[\text{Ag}]) \quad (3)$$

Assays utilizing this concept have been termed "ambient analyte immunoassays" (16), fractional occupancy being independent of both sample volume and antibody concentration (see below).

All immunoassays essentially depend on measurement of the "fractional occupancy" of the sensor antibody after its reaction with analyte (see Figure 4). Techniques relying on the measurement of unoccupied antibody binding sites (from which antibody occupancy is implicitly deduced by subtraction) necessitate—for attainment of maximal sensitivity—the use of sensor antibody concentrations tending to zero; these assays

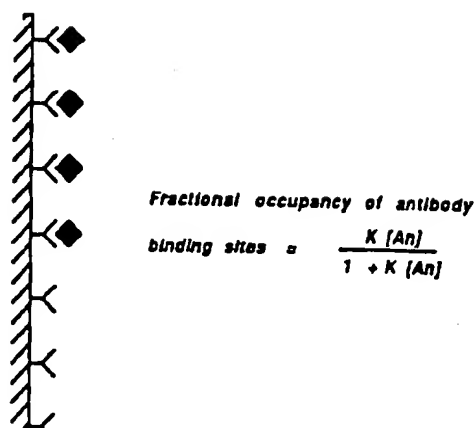


Fig. 4. The antibody binding-site occupancy principle of immunoassay

All immunoassays implicitly rely on the measurement of (fractional) binding-site occupancy by analyte

may therefore be categorized as "competitive." Conversely, techniques in which occupied sites are *directly* measured permit (in principle) the use of relatively high concentrations of sensor antibody and may be described as "noncompetitive." This difference in assay design simply reflects the proposition that, to minimize error in the measurement, it is generally undesirable to measure a small quantity by estimating the difference between two large quantities.

These concepts are illustrated in Figure 5, which portrays basic immunoassay formats currently in common use. Conventional RIA and other similar "labeled-analyte" techniques rely on measurement of *unoccupied* binding sites, generally by back-titration (either simultaneous or sequential) with labeled analyte, but anti-idiotypic antibody (reactive only with unoccupied sites on the sensor antibody) may be used for the same purpose. In the case of single-site labeled-antibody assays, the labeled antibody itself constitutes the sensor antibody; after reaction with analyte, this sensor antibody may be separated into occupied and unoccupied fractions through use of (e.g.) an immunosorbant (comprising antigen, antigen analog, or anti-idiotypic antibody linked to a solid support). If, after separation, the "signal" emitted by labeled antibody *bound* to analyte (i.e., the "occupied" fraction) is measured directly, the assay can be classed as "noncompetitive." Conversely, if one measures the labeled antibody *not bound* to analyte (i.e., that attached to the immunosorbant), then the assay is "competitive."

Two-site "sandwich" assays are clearly more complex because they rely on two antibodies and can be considered from two points of view. For our present purposes, the solid-phase antibody can be regarded as the "sensor" antibody, with the labeled antibody enabling the occupied sensor-antibody binding sites to be distinguished. Seen from this viewpoint, two-site assays may be classed as "noncompetitive."

These considerations emphasize that the differences in design distinguishing so-called competitive and non-competitive methods are essentially unrelated to which

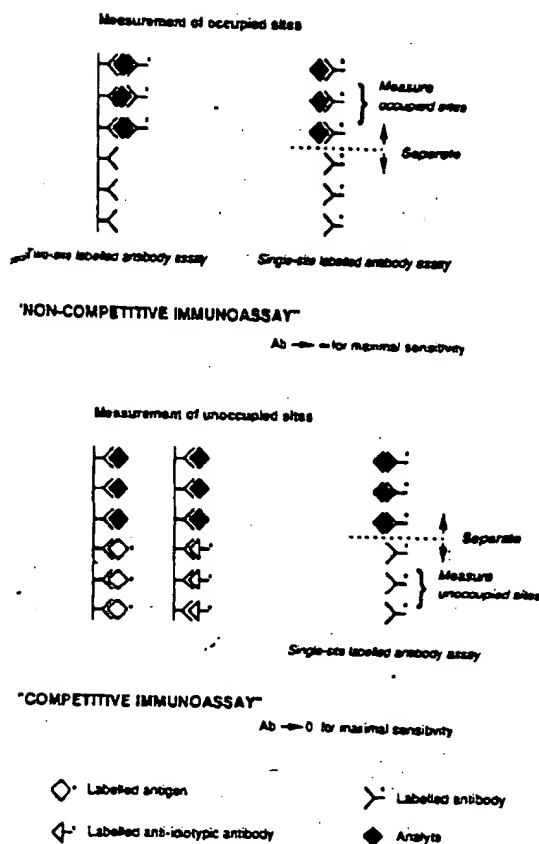


Fig. 5. Basic competitive and noncompetitive immunoassay designs

The distinction between noncompetitive and competitive immunoassays reflects the way in which antibody binding-site occupancy is observed. Labeled-antibody methods are "noncompetitive" if occupied sites of the (labeled) antibody are directly measured, but are "competitive" (lower right) when unoccupied sites are measured. Labeled-antigen (lower left) or labeled-anti-idiotypic-antibody methods (lower center) rely on measurement of sites unoccupied by analyte, and are therefore of "competitive" design

component (if any) of the reaction system is labeled. Indeed, in the case of transducer-based "immunosensors," no component is labeled; nevertheless, the design of the immunosensor will differ significantly, depending on whether a measurable signal is yielded by occupied or unoccupied antibody binding sites situated on its surface. In short, the terms "competitive" and "noncompetitive" merely reflect alternative approaches to the determination of the occupancy of antibody binding sites and lead to differences in the optimal antibody concentration required to minimize the effects of random errors arising in the determination.

Competitive and noncompetitive immunoassays can be shown to differ significantly in many of their performance characteristics, including their sensitivities. In both types of assays, both the affinity constant (K) of the antibody and the specific activity of the label are important in determining sensitivity; however, in practice, the sensitivity of competitive assays is primarily limited by the affinity constant of the antibody, whereas the specific activity of the label is more important in non-competitive systems. In both cases, the "experimental" or "manipulation" error in the measurement of the zero-dose response (R_0) [i.e., the relative error (σ_{R_0}/R_0) arising from pipetting and other operations, but not including the statistical signal measurement error per

se] is of key importance in determining "potential" assay sensitivity (i.e., the sensitivity obtained by assuming the specific activity of the label to be infinite, implying zero error in signal measurement). Thus the potential sensitivity of a competitive assay can be shown to be σ_R/KR_0 , whereas that of a noncompetitive assay is given by $R_0\sigma_{R_0}/[Ab]KR_0$, where, in the latter case, R_0 is assumed to represent the labeled antibody misclassified as bound ($[bAb]_0$), commonly referred to as "nonspecifically bound" antibody. Thus $R_0/[Ab] = f$, the fraction of labeled antibody that is nonspecifically bound, and $R_0\sigma_{R_0}/[Ab]KR_0 = f\sigma_{R_0}/KR_0$. Assuming that the relative error (σ_R/R_0) in the measurement of the zero-dose response is approximately identical for both competitive and noncompetitive assays, it is evident from this simple analysis that the potential sensitivity of noncompetitive methods is greater than that of competitive methods by the factor f , i.e., by the fraction of labeled antibody that is "nonspecifically bound." For example, if the nonspecifically bound fraction is 0.01%, a noncompetitive strategy is potentially capable of a sensitivity 10 000-fold greater than that of a competitive approach, other factors being equal.

These findings are summarized in Figure 6 (left), which shows the relationships between sensitivity (expressed in terms of molecules per milliliter) and anti-

body affinity in an optimized competitive (labeled analyte) assay. For this analysis, we assume (a) the use of a label of infinite specific activity, and (b) the use of ^{125}I as a label, the radioactivity of the samples being counted for 1 min. Computations of the theoretically optimal reagent concentrations (on which calculations represented in Figure 6 rely) were based on the further assumptions that (c) the radioactivity of the antibody-bound labeled-analyte fraction was counted and (d) the (relative) "experimental error" component in the measurement of the bound fraction (σ_b/b) was 1%. Given these assumptions, the "potential" sensitivity attainable in such an assay is σ_b/Kb , where K is the affinity constant of the antibody. [For example, if the affinity constant is 10^{12} L/mol, and σ_b/b is 0.01 (1%), maximal assay sensitivity is 10^{-14} mol/L, or $\sim 6 \times 10^6$ molecules/mL.] The additional "signal measurement error" arising in consequence of counting radioactive samples for a finite time implies a loss of assay sensitivity, as shown by the upper curve in Figure 6 (left). However, the resulting loss in sensitivity is relatively small for antibodies of affinities $<10^{12}$ L/mol, and is negligible for antibodies with affinities $<10^{11}$ L/mol. In other words, if the assayist can accept individual sample counting times of 1–5 min, little improvement in sensitivity is gained by using alternative labels of higher specific activities than ^{125}I . However, similar considerations suggest that radioisotopic labels of much lower specific activity than ^{125}I (e.g., ^3H) may limit the sensitivities of the assays (such as steroid assays) in which they are used, notwithstanding the use of relatively long sample counting times.

The other main conclusions stemming from such analysis are the importance of both minimizing "manipulation" errors and using antibodies of high binding affinity. For example, an increase in σ_b/b to 3% implies an approximate threefold loss in sensitivity, notwithstanding the fact that an assay reoptimized in response to the deterioration in operator skill that these numbers imply would utilize less antibody and labeled analyte, thereby partially offsetting the consequences of poor pipetting. But the most important conclusion emerging from the analysis is the near impossibility, in practice, of achieving immunoassay sensitivities better than about 10^7 molecules/mL by using a competitive approach, irrespective of the nature of the label used, if one assumes an upper limit to antibody binding affinities on the order of 10^{12} L/mol.

The results of a similar analysis of the sensitivity limitations applying to noncompetitive (two-site) assays (15) are illustrated in Figure 6 (right). Two sets of curves are portrayed here, corresponding to the assumptions of 1% and 0.01% nonspecific binding of labeled antibody to the capture-antibody substrate. Such analysis likewise yields important conclusions relevant to assay design, e.g., the crucial importance of reducing nonspecific binding of labeled antibody to an absolute minimum. Furthermore, if nonspecific binding is reduced to $\sim 0.01\%$, just as high sensitivity is achievable

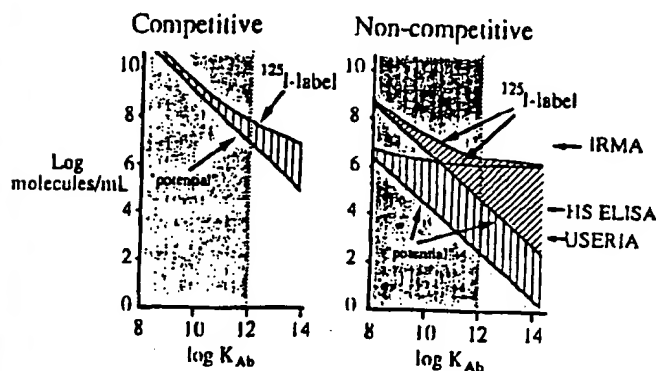


Fig. 6. Theoretically predicted sensitivities of competitive and noncompetitive immunoassay methods (represented by the SD of zero analyte measurements, expressed as molecules/mL) plotted as a function of antibody affinity (K)

Note: in noncompetitive sandwich assays, the antibody affinity referred to is that of the labeled antibody. In the competitive assays, calculations are based on the assumption that the experimental error (CV) incurred in the measurement of the assay response (e.g., fraction of labeled antigen bound) is 1%. The "potential sensitivity" curve assumes the use of a label of infinite specific activity, implying that the error in the measurement of the label per se is zero. The ^{125}I -label curve indicates the loss in sensitivity arising from the statistical error incurred in counting ^{125}I disintegrations for a finite counting time. Note that, if using antibodies with an affinity $<10^{12}$ L/mol (the maximum achieved in practice), little increase in sensitivity can be achieved by using labels of higher specific activity than ^{125}I . For noncompetitive assays, the potential sensitivity curves shown relate to values of nonspecific binding of labeled antibody of 1% (upper curves) and 0.01% (lower curves), and emphasize the improvement in sensitivity potentially attainable by minimizing nonspecific binding. The corresponding ^{125}I -label curves demonstrate the much greater loss in sensitivity (compared with that potentially attainable) when a radiolabel is used, and the special advantages of nonisotopic labels of higher specific activity in noncompetitive assay designs (particularly if nonspecific binding is reduced to 0.1% or less). Arrows indicate assay sensitivities reported for noncompetitive immunoassays based on ^{125}I (IRMA), and enzymes relying on fluorogenic (HS-ELISA) (28) and radioactive (USERIA) (29) substrates. These conclusions underlay the original development (19, 20) of time-resolved fluorimmunoassay (DELFA), the first nonisotopic "ultra-sensitive" immunoas-

by using an antibody of $K = 10^8$ L/mol in an optimized noncompetitive assay design as by using an antibody of $K = 10^{12}$ L/mol in a competitive method. One of the most important conclusions is that the sensitivities potentially attainable with high-affinity antibodies ($K > 10^{10}$ L/mol) are beyond the reach of radioisotopically based methods, which (because of the relatively low specific activities of isotopes such as ^{125}I) are limited in practice to sensitivities of the order of 10^6 – 10^7 molecules/mL or more. In short, although, under certain circumstances, noncompetitive IRMAS may be somewhat more sensitive than corresponding RIA techniques (assuming the use of the same antibody in each methodology), the potential advantages (*vis-à-vis* sensitivity) of the noncompetitive approach can be realized only by using nonisotopic labels of much higher specific activity than ^{125}I . The superiority of such labels is most apparent when they are combined with high-affinity antibodies; however, Figure 6 demonstrates that, even with use of antibodies with affinities of about 10^8 – 10^9 L/mol, nonisotopic labels may yield a substantial improvement in sensitivity.

These theoretical conclusions, together with the publication by Köhler and Milstein (18) of methods of *in vitro* production of monoclonal antibodies (1), constituted the basis of my laboratory's collaborative development (initiated around 1976) with the instrument manufacturer LKB/Wallac of the time-resolved fluorometric immunoassay methodology now known as DELFIA (19, 20). This methodology was the first "ultra-sensitive" nonisotopic immunoassay methodology to be developed. The same basic approach has subsequently been adopted by many other manufacturers, using a variety of high-specific activity labels (Table 1).

Against this background, let us now turn to the development of highly sensitive, miniaturized "microspot" immunoassays and multianalyte assay systems.

Antibody "Microspot" Immunoassay: Basic Concepts and Theory

Ambient Analyte Immunoassay

Particular attention has been drawn above to the specious notion that an antibody concentration approximating $0.5/K$ is required to maximize the sensitivity of conventional labeled-antigen assays. This proposition is implicitly overturned by the development of "microspot" immunoassays, which we expect to provide the basis of a new generation of binding assay methods. But before

discussing this methodology in detail, another basic analytical concept must be examined.

The recognition that all immunoassays essentially rely on measurement of antibody occupancy leads to a potentially important type of assay, ambient analyte immunoassay (16). This name is intended to describe assay systems that, unlike conventional methods, measure the analyte concentration in the medium to which an antibody is exposed, being independent both of sample volume and of the amount of antibody present. The possibility of developing such assays follows from the Law of Mass Action, which leads to the following equation, representing the fractional occupancy (F) by analyte of antibody binding sites (at equilibrium):

$$F^2 - F\{([1/(\underline{Ab})] + ([\underline{An}]/\underline{Ab}) + 1) + [\underline{An}]/\underline{Ab}\} = 0 \quad (4)$$

where $[\underline{An}]$ = analyte concentration, $[\underline{Ab}]$ = antibody concentration (both in units of $1/K$).¹

From this equation it may readily be shown that, for antibody concentrations approaching 0, $F \approx [\underline{An}]/(1 + [\underline{An}])$. This conclusion is illustrated in Figure 7, in which the fractional occupancy of ("monospecific" or "monoclonal") antibody binding sites in the presence of various analyte concentrations is plotted against antibody concentration. When an antibody concentration of less than (say) $0.01/K$ (the antibody preferably, but not essentially, being coupled to a solid support) is exposed to an analyte-containing medium, the resulting (fractional) occupancy of antibody binding sites solely reflects the ambient concentration of analyte¹ and is independent of the total amount of antibody in the system. (If, for example, $K = 10^{11}$ L/mol, an antibody binding-site concentration of $0.01/K$ represents 0.01×10^{-11} mol/L, or 6.02×10^7 binding sites/mL.) Analyte binding by antibody causes depletion of (unbound) analyte in the medium but, because the amount bound is small, the resulting reduction in the ambient concentration of analyte is insignificant. For example, if the concentration of binding sites of the sensor antibodies is $< 0.01/K$, analyte depletion in the medium is invariably $< 1\%$, and the system is therefore effectively independent

¹ Expression of reagent concentrations in terms of $1/K$ units has the effect of generalizing the graphical representation of binding assay data. The terms $[\underline{Ab}]$ and $[\underline{An}]$ are underlined to indicate that this convention has been adhered to in deriving equation 4. They do not refer to molar concentrations and are not interchangeable with (\underline{Ab}) and (\underline{An}) . For example, if the antibody possesses an affinity (constant) for analyte of 10^{11} L/mol, a concentration of 10^{-11} mol/L (represented in units of $1/K$) is 1 (dimensionless) unit. Thus, fractional occupancy curves based on equation 4 are identical for all antibodies if this way of expressing antibody concentration is adopted: i.e., curves relating F to analyte concentration will be identical for systems using 10^{-11} mol/L concentrations of an antibody with an affinity of 10^{11} L/mol, 10^{-10} mol/L of an antibody with an affinity of 10^{10} L/mol, 10^{-9} mol/L of an antibody with an affinity of 10^9 L/mol, etc. (provided the analyte concentration is expressed in the same manner).

² The term "ambient" is used to indicate that antibody occupancy reflects the analyte concentration to which antibody binding sites are exposed, not the amount of analyte in the incubation tube; i.e., the system is independent of sample volume.

Table 1. Detection Limits According to Type of Label

Label	Specific activity
¹²⁵ I	1 detectable event per second per 7.5×10^6 labeled molecules
Enzyme label	Determined by enzyme "amplification factor" and detectability of reaction product
Chemiluminescent label	1 detectable event per labeled molecule
Fluorescent label	Many detectable events per labeled molecule

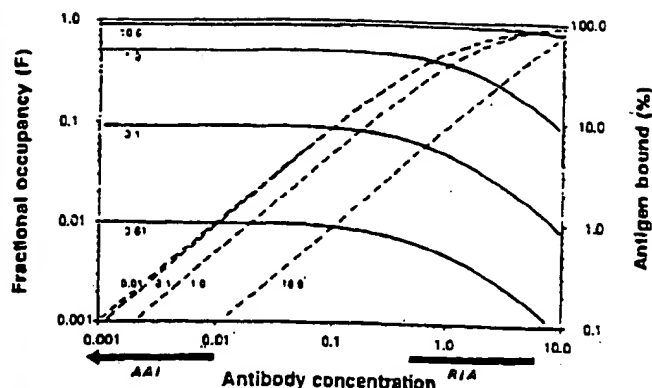


Fig. 7. Fractional antibody binding-site occupancy (F , see equation 4) plotted as a function of antibody binding-site concentration for different values of analyte (antigen) concentration (—), and the percentage binding (b) of analyte to antibody (right-hand ordinate; - -)

All concentrations are expressed in units of $1/K$. Note that for antibody concentrations $< 0.01/K$ (approximately), the percentage binding of analyte is $< 1\%$ for all analyte concentrations, and fractional binding-site occupancy is essentially unaffected by variations in antibody concentration extending over several orders of magnitude, being governed solely by antigen concentration (ambient analyte immunoassay). Note that radioimmunoassays and other "competitive" immunoassays are conventionally designed to use antibody concentrations approximating $0.5/K$ – $1/K$ or more (implying binding of analyte concentrations tending to zero (b_0) $> 30\%$), in accordance with the precepts of Yellow and Berson (1, 3).

dent of sample volume.

These conclusions lead to two further concepts. First, the antibody may be confined to a "microspot" on a solid support, such that the total number of antibody binding sites within the microspot is $< v/K \times 10^{-5} \times N$, where v = the sample volume to which the microspot is exposed (in milliliters) and N = Avogadro's number (6×10^{23}). For example, if $v = 1$ and $K = 10^{12}$ L/mol, then the

maximum number of binding sites that will cause negligible disturbance ($< 1\%$) to the ambient concentration of analyte is 6×10^6 , this number being greater for lower-affinity antibodies. Furthermore, the perception that the ratio of occupied (or unoccupied) sites to total binding sites is solely dependent on the ambient concentration of analyte leads to the concept of a dual-label, "ratiometric," microspot immunoassay.

Dual-Label Microspot Immunoassay

After exposure of a microspot of antibody (located on a suitable probe) to an analyte-containing fluid (see Figure 8, left), the probe may be removed and exposed to a solution containing a high concentration of a "developing" antibody directed against either a second epitope (i.e., the occupied site) on the analyte molecule if the molecule is large, or against unoccupied binding sites on the antibody in the case of small analyte molecules (Figure 8, right). The fractional occupancy of the sensor antibody may thus be estimated by measuring the ratio of sensor and developing antibodies that form the dual-antibody "couplets." This can be readily achieved by labeling the sensor and the developing antibodies with different labels, e.g., a pair of radioactive, enzyme, or chemiluminescent markers (or even labels of entirely different nature). Fluorescent labels are potentially particularly useful in this context because, by the use of optical scanning techniques (Figure 9), they permit the scanning of arrays of antibody "microspots" distributed over a surface (each microspot directed against a different analyte), so that multiple analyte assays may be performed simultaneously on the same sample. Several

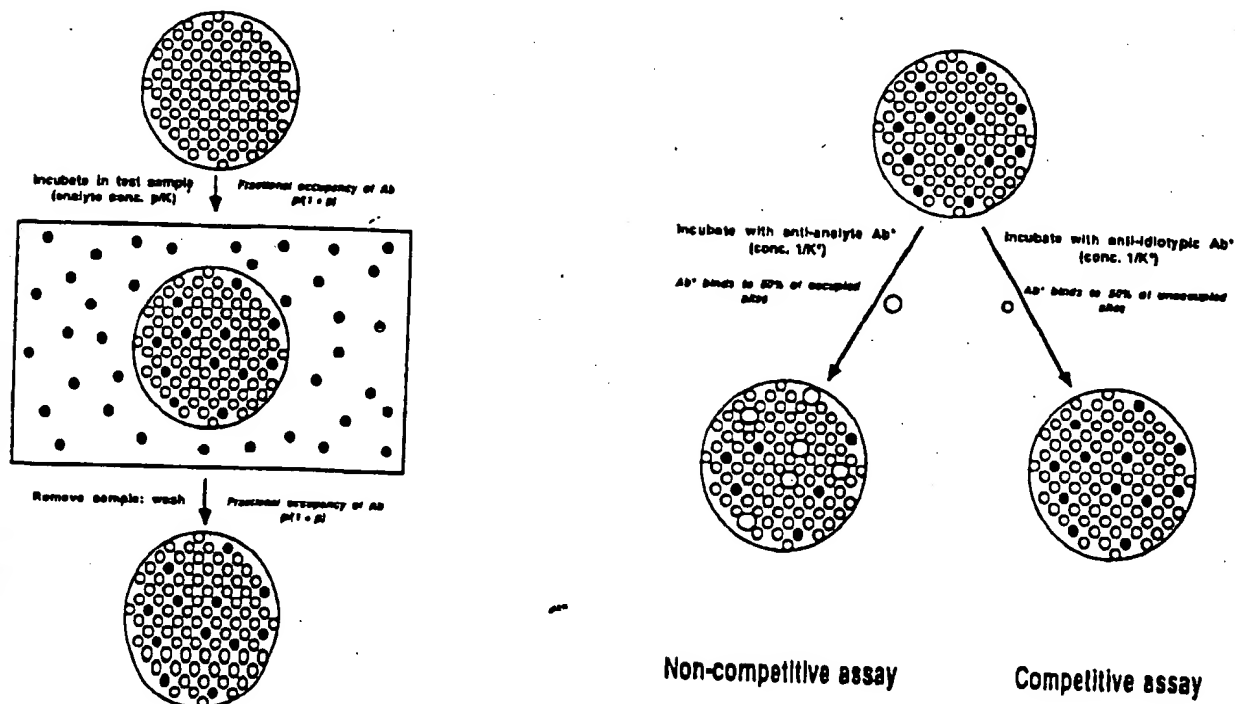


Fig. 8. Microspot immunoassay: (left) first incubation, with the fractional occupancy of antibody binding sites reflecting the analyte concentration to which the microspot has been exposed; (right) second incubation, in which the microspot is exposed to a second "developing" antibody reactive with either occupied sites (noncompetitive assay), or unoccupied sites (competitive assay). In the second incubation, a concentration of developing antibody has been selected such that only 50% of the occupied or unoccupied sites is identified

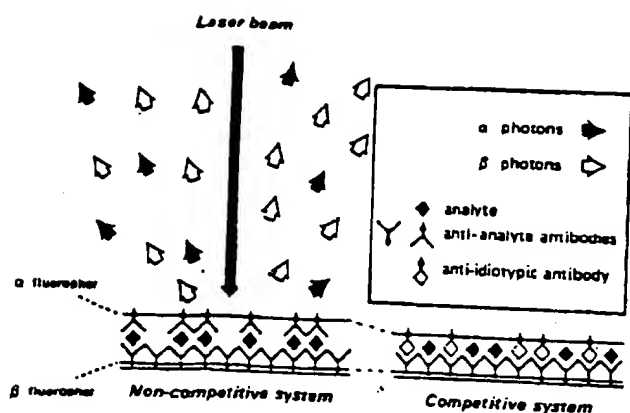


Fig. 9. Basic principle of dual-label, ambient analyte immunoassay relying on fluorescent-labeled antibodies

The ratio of α and β fluorescent photons emitted reflects the value of F (see Fig. 7) and depends solely on the analyte concentration to which the probe has been exposed. The ratio is unaffected by the amount or distribution of antibody coated (as a monomolecular layer) onto the probe surface

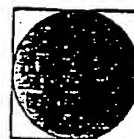
advantages stem from adopting a dual fluorescence measurement. For example, neither the amount nor the distribution of the sensor antibody within the detector's field of view is important, because the ratio of the emitted fluorescent signals is unaffected. Likewise, fluctuations in the intensity of the incident (exciting) light beam are apt to be of little significance. These advantages are additional to the basic benefit stemming from this approach, i.e., that the necessity of ensuring constancy of the amount of sensor antibody used in the assay system is removed.

Microspot Immunoassay Sensitivity

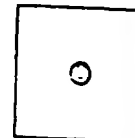
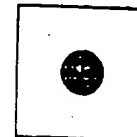
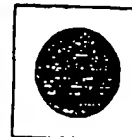
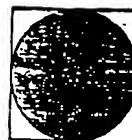
Because the microspot immunoassay methodology challenges concepts that have dominated immunoassay design theory in the past two to three decades, consideration of the potential sensitivity attainable by this approach is obviously of primary importance. The proposition that microspot assays may be at least as sensitive as conventional systems that rely on far larger amounts of antibody may readily be demonstrated by consideration of a model system. Let us postulate that sensor antibody molecules are attached to the surface of a solid support such that their binding sites remain exposed to the analyte, and that their affinity for the analyte is thereby unchanged. (The antibody concentration in the system—the number of binding sites on the support divided by the incubation volume—is unaffected by such attachment, and antibody occupancy by analyte at equilibrium will be identical to that occurring if the antibody is distributed uniformly throughout the incubation mixture.) Let us also suppose that the antibody molecules exist as a uniform monolayer of maximal surface density on the support and (to simplify discussion) are unlabeled. Then a change in the concentration of sensor antibody implies a corresponding change in the surface area over which the antibody is distributed. If, for example, the antibody affinity constant is 10^{11} L/mol, the total incubation volume is 1 mL, and the antibody surface density is 6000 binding sites/ μm^2 , then

a surface area of $10^5 \mu\text{m}^2$ (i.e., 0.1 mm^2) accommodates antibody binding sites corresponding to a concentration of $0.1/K$; an area of 0.01 mm^2 corresponds to a concentration of $0.01/K$, etc. Let us further postulate that, after exposure of the sensor antibodies to a medium containing analyte at a concentration of $0.01/K$ (i.e., 6×10^7 molecules/mL), we measure "noncompetitively" the resulting antibody occupancy (e.g., by exposure to a second, labeled, "developing" antibody directed against the analyte, forming a typical antibody sandwich). Finally, let us suppose that all occupied sites react with the developing antibody, with the latter also binding "non-specifically" to the solid support itself at a surface density of 1 molecule/ μm^2 .

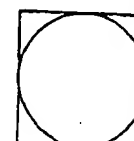
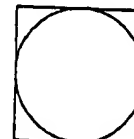
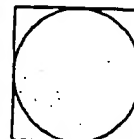
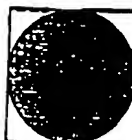
We may now consider the effects of a progressive reduction of the antibody-coated surface area from (e.g.) 1 mm^2 (effective antibody concentration $1/K$) through 0.1 mm^2 ($0.1/K$) to 0.01 mm^2 ($0.01/K$) and below. From equation 4, the value of F for the 1 mm^2 area is 4.98×10^{-3} . Thus at equilibrium the number of analyte and labeled antibody molecules specifically bound to the area is 2.99×10^7 (i.e., about 50% of the total analyte molecules present), whereas the number of labeled antibody molecules nonspecifically bound is 10^6 . Thus, assuming the field of view of the detecting instrument is restricted to the area on which the sensor antibody is deposited (see Figure 10a), and (provisionally) assuming the background (or "noise") of the instrument itself to be zero (i.e., the only source of background is the non-



a. Field of view decreases; area of antibody deposition decreases: S/B rises



b. Field of view constant; area of antibody deposition decreases: S/B falls



c. Field of view constant; density of antibody deposition decreases: S/B falls

Fig. 10. "Capture" antibody (CAb) is assumed coated on circular (shaded) areas; the field of view of the signal-measuring instrument is represented by square (unshaded) areas

(a) Reduction of both the area of deposition of CAb and the field of view results in an increase in the signal/noise (S/B) ratio. If the CAb is reduced either by reducing the antibody coated area (b) or the density of antibody coating (c) while the field of view remains unchanged, S/B falls

specifically-bound labeled antibody within the instrument's field of view), the signal/noise ratio observed for the 1 mm² area is ~30. Similarly, the value of F for a 0.1 mm² area is 9.02×10^{-3} , the number of labeled antibody molecules specifically bound to the area is 5.41×10^6 , the number nonspecifically bound is 10^6 , and the signal/noise ratio is ~54. Likewise, the signal/noise ratio for a 0.01 mm² area can be shown to be ~59. In short, the signal/noise ratio increases as the antibody-coated surface area is decreased, approaching a maximal (plateau) value of 60 as the area coated with sensor antibody falls below 0.01 mm² and tends toward zero.

If, however, a reduction in the antibody-coated area were not accompanied by a corresponding reduction in the detecting instrument's field of view, the resulting reduction in "signal" would not lead to a corresponding decrease in the background generated by nonspecifically-bound developing antibody (Figure 10b). Therefore, although reduction in the coated area would increase the fractional occupancy of the sensor antibody, the signal/noise ratio might either remain constant or fall. In these circumstances it might be advantageous to increase the coated area. Similarly, if the surface density of sensor antibody were decreased (the coated area being held constant), similar conclusions would be reached (Figure 10c).

Likewise, if the background signal generated within the detecting instrument itself (e.g., from the photocathode of a photomultiplier tube used to detect photons emitted from the antibody-coated area) were not zero, and remained constant regardless of the instrument's field of view, then a maximum signal/noise ratio would also be attained at some optimal value of the antibody-coated area, below which the ratio would fall. Because, however, one can generally reduce the size of the detector (and hence the detector-generated background) at the same rate as the size of the signal-emitting area, there is no reason—in principle—for the signal/noise ratio to diminish as the antibody-coated area is progressively reduced toward zero. Thus if we accept the signal/noise ratio as indicative of the precision of the measurement of antibody occupancy (and hence of assay sensitivity), these considerations suggest that it is advantageous to reduce the antibody-coated surface area (and, concomitantly, the sensor-antibody concentration) toward zero, although little advantage is likely to accrue from reducing the area below 0.01 mm² (and thus the antibody concentration below $0.01/K$).

Were the microspot area indeed reduced to zero, both signal and noise would likewise also fall to zero (the ratio between them nevertheless remaining essentially constant), implying that no signal of any kind would, in the limit, be recorded. In practice, other statistical factors come into play when the number of individual events (e.g., photons) observed by a detecting instrument is very low, thus prohibiting a reduction of the sensor antibody concentration to zero. The point at which the reduction in the antibody-coated area causes the detectable signal to be lost sufficiently to affect the

precision of the measurement of antibody occupancy depends clearly on the specific activity of the labeled antibody used to measure the occupied binding sites: the higher the specific activity, the smaller the permissible area. Thus, given labels of very high specific activity, one can envision circumstances in which, even in a "noncompetitive" system, the optimal concentration of sensor antibody may be exceedingly low. A more general conclusion is that a variety of factors, including the characteristics of the instruments used for measuring the labeled antibody (or labeled analyte), influence immunoassay design, implying, among other things, the virtual impossibility of formulating general rules regarding this. For example, reagent concentrations that are optimal for isotopically labeled reagents used with a conventional radioisotope counter (possessing a fixed background dependent on its basic construction) are likely to be entirely different when very high-specific-activity labels are used and one has the freedom to tailor the measuring instrument to samples of any size. In short, certain conclusions based on experience of RIA and IRMA techniques may prove misleading when applied to nonisotopic methodologies, and should be viewed with caution.

A more detailed theoretical consideration of (noncompetitive) microspot immunoassay sensitivity (21) suggests that

$$C_{\min} = D^*_{\min} \times [(6 \times 10^{20})(1 + [Ab^*])]/DK[Ab^*] \quad (5)$$

where D = surface density (binding sites/ μm^2) of sensor antibody, K = sensor antibody affinity (L/mol), $[Ab^*]$ = concentration of labeled antibody in developing solution (expressed in units of $1/K^*$, where K^* = labeled antibody affinity), D^*_{\min} = minimum detectable surface density of labeled antibody (molecules/ μm^2), and C_{\min} = assay detection limit (molecules/mL). For example, if $[Ab^*] = 1$, $D = 10^6$ molecules/ μm^2 , $K = 10^{11}$ L/mol, and $D^*_{\min} = 20$ molecules/ μm^2 , then $C_{\min} = 2.4 \times 10^6$ molecules/mL = 4×10^{-16} mol/L and the fractional occupancy of the binding sites of the sensor antibody by the minimum detectable concentration of analyte is 0.04%. Figure 11 shows the theoretical assay sensitivities attainable with use of sensor antibodies of various affinities, plotted as a function of D^*_{\min} .

A similar theoretical analysis of competitive microspot immunoassay indicates that potential sensitivities are essentially identical to those attainable with conventional competitive methodologies. In summary, the above considerations indicate that the attainment of high microspot assay sensitivity requires close packing of molecules of sensor antibodies within the microspot area, combined with the use of an instrument capable of accurately measuring very low surface densities of developing antibodies. They also suggest that (a) microspot assay sensitivities considerably higher than those obtainable by conventional isotopically based immunoassays are achievable, and (b) if labels of very high specific activity are available, the sensitivities yielded

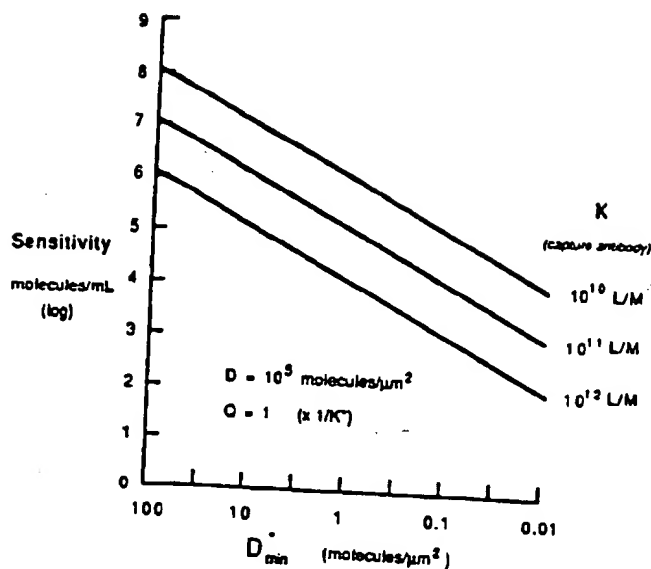


Fig. 11. Theoretically predicted sensitivity of noncompetitive microspot immunoassay plotted as a function of the minimum developing antibody density detectable within the microspot area

Postulated values of capture antibody surface density are 10^6 molecules/ μm^2 , and of developing antibody concentration are $1/K$. Currently available instruments permit detection of between 10 and 1 molecules of fluorescein-labeled antibody per micrometer.²

by microspot assays are unlikely to be inferior and (depending on the characteristics of the measuring instruments used) could be superior to the sensitivities achievable in macroscopic assays of conventional design.

Finally, we briefly address a further question occasionally raised in this context, i.e., the kinetic characteristics of microspot assays. Two points should be made regarding this issue. First, the smaller the microspot of sensing antibody, the lower the diffusion constraints on the velocity of the antibody/analyte binding reaction, so that at the limit (i.e., when the amount of antibody situated within the microspot area approaches zero) the kinetics of the reaction approximate those observed in a homogeneous liquid-phase system. Second, although the effective concentration of sensor antibody in the incubation medium is exceedingly low, the fractional rate at which sensor antibody binding sites within the microspot become occupied is invariably greater in this circumstance than when a relatively high concentration of antibody is used, as in conventional assays, particularly those of noncompetitive design. In other words, bearing in mind the relationship between fractional occupancy of sensor antibody and the signal/noise ratio discussed above, it is readily demonstrable that the rate at which the ratio rises is greatest when the microspot area (and the antibody contained within it) is least. Thus, given instrumentation whose field of view is restricted to the microspot area, the highest signal/noise ratio will be observed (after any selected incubation period) when the concentration of sensor antibody in the system is $<0.01/K$. In short, contrary perhaps to superficial impression, and to the generally accepted belief that short immunoassay incubation times require the use of very large amounts of antibody, the antibody microspot ap-

proach provides the basis of assays potentially more rapid than any currently available.

Microspot Immunoassay: Some Practical Considerations

Although various high-specific-activity antibody labels are potentially usable in this context, our preliminary studies have relied on the use of conventional fluorophors. The simultaneous measurement of dual fluorescences from small areas is, of course, well established, and the availability of improved instrumentation (e.g., the laser scanning confocal microscope), albeit not specifically designed for the present purpose, has been useful in demonstrating the feasibility of the microspot approach.

In laser scanning confocal fluorescence microscopes, a small area of the specimen is illuminated by a focused laser beam, the fluorescence photons emitted from this area being focused in turn onto a detector, typically a low-dark-current photomultiplier (22, 23). At the "confocal" point, the projection of the illumination pinhole and the back-projection of the detector pinhole coincide (Figure 12). Fluorescence photons emitted at other points thus possess a low probability of reaching the detector. Such systems contrast with conventional epi-fluorescence microscopes, in which the specimen is exposed to an essentially uniform flux of illumination, and yield much sharper images of fluorescent emitters situated in a defined plane of a tissue sample. Electrons spontaneously emitted by the photomultiplier photocathode contribute to the background signal of the instrument, and must—for highest microspot assay sensitivity—be minimized. Fortunately, the design of such instruments permits the photocathode to be very small in area, and this source of background can be expected

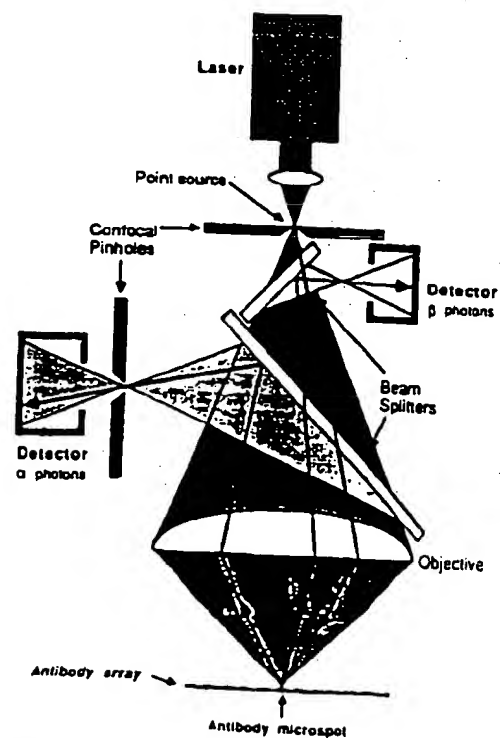


Fig. 12. Schematic diagram of a dual-channel confocal microscope

to diminish with future improvement in photomultiplier design. Other sources of background include fluorescence emitted by components in the optical system, which may not, in current instruments, have been constructed with background reduction as a prime consideration. Nevertheless, they detect with high sensitivity fluorescent signals. For example, one commercially available microscope is claimed to detect fluorescein at a density of 10 molecules/ μm^2 . Most commercially available fluorescein isothiocyanate (FITC)-labeled IgG exhibits a fluorophor/protein ratio of ~ 4 ; this implies detection limit (D^*_{min}) for antibody surface density of two or three FITC-labeled IgG molecules per micrometer². This, in turn, implies a theoretical sensitivity for a two-site immunoassay of $\sim 2\text{--}3 \times 10^5$ analyte molecules per milliliter, assuming identical parameter values as above, or $2\text{--}3 \times 10^4$ molecules/mL if the sensing antibody has an affinity of 10^{12} L/mol. Clearly, sensitivity may be increased by loading more fluorophor either directly or indirectly onto the antibody.

Our preliminary studies have relied on a less sensitive microscope, albeit one possessing facilities for dual-fluorescence measurement. Its argon laser emits two excitation lines at 488 and 514 nm. It is thus particularly efficient in exciting blue/green-emitting fluorophores such as FITC (excitation maximum 492 nm), but is less efficient in exciting fluorophores such as Texas Red (excitation maximum 596 nm). However, the ratiometric assay principle permits considerable variation in detection efficiencies of the two labels because the specific activities of the labeled antibody species forming the antibody couplets can be chosen to yield signal ratios approximating unity. Inefficiency of the argon laser in exciting Texas Red is thus not a major handicap in this context. Though this instrument relies on a conventional microscope and not on an optical system designed for this purpose (and thus implicitly less sensitive), it permits quantification of fluorescence signals generated from microspots of any selected area. Initial studies have revealed that, under conditions that are not optimal, the instrument is capable of detecting ~ 25 FITC-labeled and (or) 150 Texas Red-labeled IgG molecules per micrometer², while scanning an area of $\sim 50 \mu\text{m}^2$.

The development of microspot immunoassays has also necessitated closer scrutiny of the mechanisms involved in the coupling of antibodies to solid supports. In the present context, these should display a capacity to adsorb (in the form of a monolayer)—or to covalently link—a high surface density of antibody combined with low intrinsic-signal-generating properties (e.g., low intrinsic fluorescence), thus minimizing background. We have examined a number of candidate materials, such as polypropylene, Teflon®, cellulose and nitrocellulose membranes, microtiter plates (clear polystyrene plates; black, white, and clear polystyrene plates), glass slides and quartz optical fibers coated with 3-(amino propyl) triethoxy silane, etc., and several alternative protocols for achieving high monolayer coating densities. These

studies have exposed phenomena neither evident nor of importance when antibody binding to solid supports is examined at a macroscopic level. Provisionally, we have used white Dynatech Microfluor microtiter plates—formulated for the detection of low fluorescence signals, and yielding high signal/noise ratios and high coating densities of functional antibodies ($\sim 5 \times 10^4$ IgG molecules/ μm^2)—for assay development, although such plates are not ideal. Indeed, deficiencies in the antibody-deposition methods used constitute the principal source of imprecision in assay results and the limitation in sensitivity that this implies. Clearly, this represents an area for further study and refinement of current coating techniques.

Notwithstanding the limitations of present instrumentation (which, among other things, does not permit the use of time-resolving techniques to distinguish two individual fluorescence signals either from each other or from background fluorescence) and the crudeness of present methods for coupling antibodies onto small areas, we have verified the theoretical concepts outlined above by comparing the performance of several assays when constructed in microspot format and when conventionally designed. Although unoptimized, ratiometric microspot assays have yielded sensitivity values closely approaching those of conventional optimized IRMA. As an example, the results of a ratiometric assay system for thyrotropin, with use of Texas Red- and FITC-labeled antibodies, are shown in Figure 13. Bearing in mind the well-known limitations of these and other "conventional" fluorophores when used as immunoassay reagent labels, such results are encouraging, although further work is clearly required to achieve the considerably greater sensitivity theoretically predicted with use of improved fluorophores, better antibody-microspotting techniques, and purpose-built (time-resolving) instrumentation.

The finding that highly sensitive immunoassays can be performed with far smaller amounts of antibody than

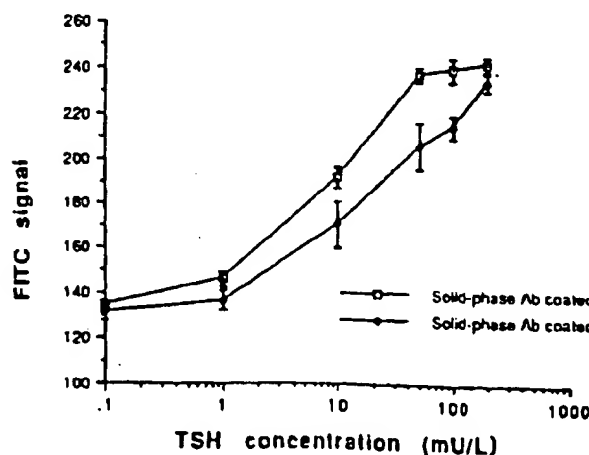


Fig. 13. Response curve in a dual-labeled microspot ratiometric assay of thyrotropin (TSH) with Texas Red-labeled solid-phase capture antibody and a developing antibody labeled with biotin/FITC-avidin

The FITC/Texas Red ratio for each microspot was measured with a scanning confocal microscope, and plotted as a function of TSH concentration in milli-int. units/L

are currently used conventionally permits in turn the construction of antibody microspot arrays enabling, in principle, the simultaneous measurement of thousands of different substances in 1-mL samples. In collaboration with investigators at the Centre for Applied Microbiological Research, Porton Down, U.K., we are presently developing various techniques for the creation of such arrays. Indeed, similar technologies have recently been used for the parallel synthesis of several different polypeptides, these enabling 10 000-microspot arrays to be constructed on silica chips approximating 1 cm^2 (24). Although arrays of this capacity are unlikely to ever be required for conventional diagnostic purposes, we can anticipate that the ability to simultaneously measure many substances in the same sample will have revolutionary consequences in medicine and other similar areas. In addition, such techniques may ultimately permit the individual analysis of the multiple isoforms of certain "heterogeneous" analytes (e.g., the glycoprotein hormones), such molecular heterogeneity currently presenting a major obstacle to the standardization and interpretation of many immunological measurements (25). Moreover, although these concepts have been illustrated in an immunoassay context, they are clearly applicable to all "binding assays," including those relying on the use of DNA probes, hormone receptors, etc. For example, labeled lectins that are specific in their reactions with the sugar residues in the oligosaccharide chains of glycoprotein molecules may be used, together with specific antibodies, to impart additional "structural specificity" to sandwich assays (26, 27), possibly overcoming the limitations of antibodies per se in regard to differentiation of the glycosylation variants of the glycoprotein hormones.

Summary and Conclusion

Because of past confusion regarding the concepts of precision, sensitivity, accuracy, etc., several erroneous concepts have become incorporated within currently accepted rules of immunoassay design. In particular, much higher antibody concentrations are customarily used than are necessary to achieve very high assay sensitivity, provided that certain measurement strategies are adhered to. In this presentation, we have attempted to show that, in principle, the highest assay sensitivities are obtained by confining a small number of sensor antibody molecules onto a very small area in the form of a microspot and measuring their occupancy by an analyte, by using very high-specific-activity "developing" antibody probes, thereby maximizing the signal/noise ratio in the determination of sensor antibody occupancy. This observation, which contradicts currently accepted immunoassay design theory, in turn makes possible the measurement of an unlimited number of different analytes on a chip of very small surface area through the use of, e.g., laser scanning techniques closely analogous to those used in compact disk techniques of sound recording. Extensive experimental studies in this area, albeit conducted with relatively crude techniques and instrumentation not specifically de-

signed for these purposes, and therefore not reported in detail here, have demonstrated the feasibility of the miniaturized antibody microspot approach and the validity of the general concepts on which it is based. We are therefore confident that this represents the basis of a next-generation technology that is likely to have a revolutionary impact on all fields involving the use of binding assays.

References

1. Yalow RS, Berson SA. General principles of radioimmunoassay. In: Hayes RL, Goswitz FA, Murphy BEP, eds. *Radioisotopes in medicine: in vitro studies*. Oak Ridge, TN: US Atomic Energy Commission, 1968:7-39.
2. Ekins RP, Newman B, O'Riordan JH. *Ibid.*: 59-100.
3. Berson SA, Yalow RS. Measurement of hormones—radioimmunoassay. In: Berson SA, Yalow RS, eds. *Methods in investigative and diagnostic endocrinology*, Vol. 2A. Amsterdam: North Holland/Elsevier, 1973:84-135.
4. Ekins R, Newman B. Theoretical aspects of saturation analysis. In: Diczfalussy E, Diczfalussy A, eds. *Steroid assay by protein binding*. Karolinska symposium on research methods in reproductive endocrinology. Stockholm: WHO/Karolinska Sjukhuset, 1970:11-30.
5. Ekins RP. Limitations of specific activity. In: Margoulies M, ed. *Protein and polypeptide hormones, Part 3 (Discussions)*. Amsterdam: Excerpta Medica, 1968:612-6, et seq.; Ekins RP. Concentrations of tracer and antiserum, time and temperature of incubation, volume of incubation. *Ibid.*: 672-82.
6. Yalow RS, Berson SA. Immunoassay of endogenous plasma insulin in man. *J Clin Invest* 1960;39:1157.
7. Ekins RP. The estimation of thyroxine in human plasma by an electrophoretic technique. *Clin Chim Acta* 1960;5:453-9.
8. Barakat RM, Ekins RP. Assay of vitamin B₁₂ in blood—a simple method. *Lancet* 1961;ii:25-6.
9. Wide L, Bennich H, Johansson SGO. Diagnosis of allergy by an in-vitro test for allergen antibodies. *Lancet* 1967;ii:1105-7.
10. Miles LEH, Hales CN. Labeled antibodies and immunological assay systems. *Nature (London)* 1968;219:186-9.
11. Keston AS, Udenfriend S, Cannan RK. Micro-analysis of mixtures (amino acids) in the form of isotopic derivatives. *J Am Chem Soc* 1946;68:1390.
12. Avivi P, Simpson SA, Tait JF, Whitehead JK. The use of ³H and ¹⁴C-labeled acetic anhydride as analytical reagents in microbiochemistry. In: Johnston JE, Faires RA, Millett RJ, eds. *Radioisotope conference*, London: Butterworths, 1954:313-23.
13. Miles LEH, Hales CN. An immunoradiometric assay of insulin. *Op. cit.* (ref. 5), Part 1:61-70.
14. Rodbard D, Weiss GH. Mathematical theory of immunometric (labeled antibody) assay. *Anal Biochem* 1973;52:10-44.
15. Jackson TM, Marshall NJ, Ekins RP. Optimisation of immunoradiometric assays. In: Hunter WM, Corrie JET, eds. *Immunoassays for clinical chemistry*. Edinburgh: Churchill Livingstone, 1983:557-75.
16. Ekins RP. Measurement of analyte concentration. British patent no. 8 224 600, 1983.
17. Wide L. Solid-phase antigen-antibody systems. In: Hunter WM, Kirkham KE, eds. *Radioimmunoassay methods*. Edinburgh: Churchill Livingstone, 1971:405-12.
18. Köhler G, Milstein C. Continuous culture of fused cells secreting specific antibody of predefined specificity. *Nature (London)* 1975;256:495-7.
19. Marshall NJ, Dakubu S, Jackson T, Ekins RP. Pulsed light, time resolved fluoroimmunoassay. In: Albertini A, Ekins RP, eds. *Monoclonal antibodies and developments in immunoassay*. Amsterdam: Elsevier/North Holland, 1981:101-8.
20. Soini E, Lövgren T. Time-resolved fluorescence of lanthanide probes and applications in biotechnology [Review]. *Crit Rev Anal Chem* 1987;18:105-54.
21. Ekins RP, Chu F, Biggart E. The development of microspot, multi-analyte radiometric immunoassay using dual fluorescent-labeled antibodies. *Anal Chim Acta* 1990;227:73-96.
22. White JG, Amos WB, Fordham M. An evaluation of confocal versus conventional imaging of biological structures by fluores-

cence light microscopy. *J Cell Biol* 1987;105:41-8.

23. Ploem JS. New instrumentation for sensitive image analysis of fluorescence in cells and tissues. In: Tayer DL, Waggoner AS, Lanni F, Murphy R, Birge R, eds. *Applications of fluorescence in the biological sciences*. New York: Alan R Liss, 1986:289-300.

24. Fodor SPA, Read JL, Pirrung MC, et al. Light-directed, spatially addressable parallel chemical synthesis. *Science* 1991;251:767-73.

25. Ekins RP. Immunoassay standardization. In: Kallner A, Magid E, Albert W, eds. *Improvement of comparability and compatibility of laboratory assay results in life sciences. Immunoassay standardization*. Scand J Clin Lab Invest 1991;51(Suppl 205):33-46.

26. Kottgen E, Hell B, Muller C, Tauber R. Demonstration of glycosylation variants of human fibrinogen, using the new tech-

nique of glycoprotein lectin immunosorbent assay (GLIA). *Biol Chem Hoppe Seyler* 1988;369:1157-68.

27. Kinoshita N, Suzuki S, Matsuda Y, Taniguchi N. α -Fetoprotein antibody-lectin enzyme immunoassay to characterize sugar chains for the study of liver diseases. *Clin Chim Acta* 1989;179:143-52.

28. Shalev V, Greenberg GH, McAlpine PJ. Detection of at-tograms of antigen by a high sensitivity enzyme-linked immuno-sorbent assay (HS-ELISA) using a fluorogenic substrate. *J Immunol Methods* 1980;88:125.

29. Harris CC, Yolken RH, Kroken H, Hsu IC. Ultrasensitive enzymatic radioimmunoassay: application to detection of cholera toxin and rotavirus. *Proc Natl Acad Sci USA* 1979;76:5336.

Corrections

Vol 37, pp. 1447-8: In our desire for rapid publication, important errors were introduced into the following Technical Brief. The corrected version is here reproduced in its entirety, with our apologies to the authors.

Rapid Detection of 1717-1G→A Mutation in CFTR Gene by PCR-Mediated Site-Directed Mutagenesis, Laura Cremonesi,¹ Manuela Seia,² Carmelina Magnani,¹ and Maurizio Ferrari² (¹ Istituto Scientifico H.S. Raffaele, Lab. Centrale, Milano; ² Istituti Clin. di Perfezionamento, Lab. di Ricerche Clin., Milano, Italy)

Until now, among the non-ΔF508 mutations identified in the cystic fibrosis transmembrane conductance regulator (CFTR) gene by the Cystic Fibrosis (CF) Genetic Analysis Consortium, the ones most frequently seen in our population sample are the 1717-1G→A mutation (13/144 or 9% of the CF chromosomes) and the G542X mutation (16/190 or 8.4% of the CF chromosomes), both revealed by dot-blot hybridization of the polymerase chain reaction (PCR) product with allele-specific oligonucleotides (ASO) probes (1).

In an attempt to simplify the analysis of the most frequent mutations in the CFTR gene, we converted radio-labeled ASO detection into restriction endonuclease analysis of the amplified product.

A PCR-mediated site-directed mutagenesis (2, 3) to detect the G542X mutation by generating a novel *Bst*NI site in the wild-type sequence had already been suggested (4).

To detect the 1717-1G→A mutation, we designed the reverse primer (5'-CTCTGCAAACCTGGAGAGGTC-3') to contain a single-base mismatch (T→G), which could create a novel *Ava*II restriction site [G ↓ G(A/T)CC] in the amplified wild-type (WT) allele but not in the CF mutant (M) allele:

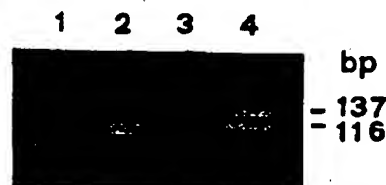
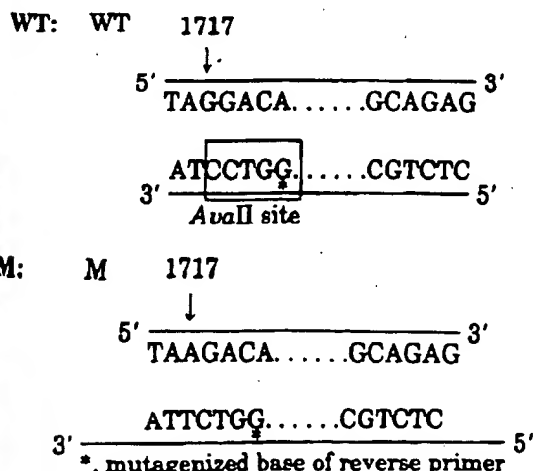


Fig. 1. Detection of the 1717-1G→A mutation by PCR

Reactions were carried out with 1 μ g of genomic DNA in a total volume of 100 μ L containing 10 mmol/L Tris-HCl (pH 8.3), 50 mmol/L KCl, 1.5 mmol/L MgCl₂, 0.1 g/L gelatin, 200 μ mol/L each of the four deoxyribonucleotide triphosphates, 2.5 units of Taq polymerase (Perkin-Elmer Cetus, Norwalk, CT), and 100 pmol of each of the primers. PCR conditions were as follows: denaturation at 94 °C for 1 min, annealing at 55 °C for 30 s, and extension at 72 °C for 1 min, for a total of 30 cycles. PCR products were digested for 2 h at 37 °C with 5 U of *Ava*II and electrophoresed on 3% agarose-1% NuSieve gel for 1 h at 50 V. Bands were made visible by staining the gel with ethidium bromide. Lane 1: *Hae*III-digested pBR322 size marker. Lane 2: normal homozygote. Lane 3: CF patient homozygous for the 1717-1G→A mutation. Lane 4: heterozygote carrier for the 1717-1G→A mutation.

For the forward primer, we used the one made available by the CF Genetic Analysis Consortium to amplify exon 11 of the CFTR gene: 5'-CAACTGTGGTTAAAGCAAT-AGTGT-3'.

Digestion by *Ava*II enzyme of the PCR product generates two fragments of 116- and 21-bp in the wild-type alleles and leaves undigested a 137-bp fragment in the mutant alleles (Figure 1).

By combined analysis for the ΔF508 mutation (5) (252/470 or 53.6% of the CF chromosomes), 1717-1G→A, and G542X, about 71% of mutations might be detected by nonisotopic analysis of the PCR product, thus allowing a faster and easier one-day procedure for carrier screening and prenatal testing.

References

- Kerem B, Zielenski J, Markiewicz D, et al. Identification of mutations in regions corresponding to the two putative nucleotide (ATP)-binding folds of the cystic fibrosis gene. *Proc Natl Acad Sci USA* 1990;87:8447-51.
- Haliassos A, Chomel JC, Baudis M, Kruh J, Kaplan JC, Kitzis A. Modification of enzymatically amplified DNA for the detection of point mutations. *Nucleic Acids Res* 1989;17:3606.
- Friedman WE, Highsmith EJ Jr, Prior TW, Perry TR, Silverman LM. Cystic fibrosis deletion mutation detected by PCR-mediated site-directed mutagenesis [Tech Brief]. *Clin Chem* 1990;36:695-6.
- Ng ISL, Pace R, Richard MV, et al. Methods for analysis of multiple cystic fibrosis mutations. *Hum Genet* (in press).
- Ferrari M, Cremonesi L. More on detection of cystic fibrosis by polymerase chain reaction [Response to Letter]. *Clin Chem* 1990;36:1702-3.

clinical chemistry

11
91

In This Issue . . .

Kornberg on Life as Chemistry

See Page 1895

Cyclosporine Monitoring

See Pages 1891, 1905

Clinical Uses of DNA Amplification

See Pages 1893, 1945, 1983

CLIA and Cholesterol Testing

See Page 1938

**American Thyroid Association
Report**

See Page 2002



INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification ⁶ : C12Q 1/68	A1	(11) International Publication Number: WO 95/21944 (43) International Publication Date: 17 August 1995 (17.08.95)
(21) International Application Number: PCT/US95/01863 (22) International Filing Date: 14 February 1995 (14.02.95) (30) Priority Data: 08/195,485 14 February 1994 (14.02.94) US (60) Parent Application or Grant (63) Related by Continuation US 08/195,485 (CIP) Filed on 14 February 1994 (14.02.94) (71) Applicant (for all designated States except US): SMITHKLINE BEECHAM CORPORATION [US/US]; Corporate Intellectual Property, UW2220, 709 Swedeland Road, P.O. Box 1539, King of Prussia, PA 19406-0939 (US). (72) Inventors; and (75) Inventors/Applicants (for US only): ROSENBERG, Martin [US/US]; 241 Mingo Road, Royersford, PA 19468 (US). DEBOUCK, Christine [BE/US]; 667 Pugh Road, Wayne, PA 19087 (US). BERGSMA, Derk [US/US]; 271 Irish Road, Berwyn, PA 19312 (US).	(74) Agents: JERVIS, Herbert, H. et al.; SmithKline Beecham Corporation, Corporate Intellectual Property, UW2220, 709 Swedeland Road, P.O. Box 1539, King of Prussia, PA 19406-0939 (US). (81) Designated States: JP, US, European patent (AT, BE, CH, DE, DK, ES, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE). Published With international search report.	
(54) Title: DIFFERENTIALLY EXPRESSED GENES IN HEALTHY AND DISEASED SUBJECTS (57) Abstract <p>The present invention involves methods and compositions for identifying genes which are differentially expressed in a normal healthy animal and an animal having a selected disease or infection, and methods for diagnosing diseases or infections characterized by the presence of those genes, despite the absence of knowledge about the gene or its function. The methods involve the use of a composition suitable for use in hybridization which consists of a solid surface on which is immobilized at pre-defined regions thereon a plurality of defined oligonucleotide/polynucleotide sequences for hybridization. Each sequence comprises a fragment of an EST isolated from an identified DNA library prepared from tissue or cell samples of a healthy animal, an animal with a selected disease or infection, and any combination thereof. Differences in hybridization patterns produced through use of this composition and the specified methods enable diagnosis of disease based on differential expression of genes of unknown function, and enable the identification of those genes and the proteins encoded thereby.</p>		

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AT	Austria	GB	United Kingdom	MR	Mauritania
AU	Australia	GE	Georgia	MW	Malawi
BB	Barbados	GN	Guinea	NE	Niger
BE	Belgium	GR	Greece	NL	Netherlands
BF	Burkina Faso	HU	Hungary	NO	Norway
BG	Bulgaria	IE	Ireland	NZ	New Zealand
BJ	Benin	IT	Italy	PL	Poland
BR	Brazil	JP	Japan	PT	Portugal
BY	Belarus	KE	Kenya	RO	Romania
CA	Canada	KG	Kyrgyzstan	RU	Russian Federation
CF	Central African Republic	KP	Democratic People's Republic of Korea	SD	Sudan
CG	Congo	KR	Republic of Korea	SE	Sweden
CH	Switzerland	KZ	Kazakhstan	SI	Slovenia
CI	Côte d'Ivoire	LI	Liechtenstein	SK	Slovakia
CM	Cameroon	LK	Sri Lanka	SN	Senegal
CN	China	LU	Luxembourg	TD	Chad
CS	Czechoslovakia	LV	Latvia	TG	Togo
CZ	Czech Republic	MC	Monaco	TJ	Tajikistan
DE	Germany	MD	Republic of Moldova	TT	Trinidad and Tobago
DK	Denmark	MG	Madagascar	UA	Ukraine
ES	Spain	ML	Mali	US	United States of America
FI	Finland	MN	Mongolia	UZ	Uzbekistan
FR	France			VN	Viet Nam
GA	Gabon				

differentially expressed genes in healthy and diseased subjects

Cross Reference to Related Applications:

5 This application is a continuation-in-part application of U.S. Serial No. 08/195,485 filed February 14, 1994, the contents of which are incorporated herein by reference.

Field of the Invention

10 The present invention relates to the use of immobilized oligonucleotide/polynucleotide or polynucleotide sequences for the identification, sequencing and characterization of genes which are implicated in disease, infection, or development and the use of such identified genes and the proteins encoded thereby in diagnosis, prognosis, therapy and drug discovery.

15

Background of the Invention

 Identification, sequencing and characterization of genes, especially human genes, is a major goal of modern scientific research. By identifying genes, determining their sequences and characterizing their biological function, it is possible to employ recombinant DNA technology to produce large quantities of valuable "gene products", e.g., proteins and peptides. Additionally, knowledge of gene sequences can provide a key to diagnosis, prognosis and treatment of a variety of disease states in plants and animals which are characterized by inappropriate expression and/or repression of selected gene(s) or by the influence of external factors, e.g., carcinogens or teratogens, on gene function. The term disease-associated genes(s) is used herein in its broadest sense to mean not only genes associated with classical inherited diseases, but also those associated with genetic predisposition to disease as well as infectious or pathogenic states resulting from gene expression by infectious agents or the effect on host cell gene expression by the presence of such a pathogen or its products. Locating disease-associated genes will permit the development of diagnostic and prognostic reagents and methods, as well as possible therapeutic regimens, and the discovery of new drugs for treating or preventing the occurrence of such diseases.

 Methods have been described for the identification of certain novel gene sequences, referred to as Expressed Sequence Tags (EST) [see, e.g., Adams et al, *Science*, 252:1651-1656 (1991); and International Patent Application No. WO93/00353, published January 7, 1993]. Conventionally, an EST is a specific cDNA polynucleotide sequence, or tag, about 150 to 400 nucleotides in length, derived from

a messenger RNA molecule by reverse transcription, which is a marker for, and component of, a human gene actually transcribed *in vivo*. However, as used herein an EST also refers to a genomic DNA fragment derived from an organism, such as a microorganism, the DNA of which lacks intron regions.

5 A variety of techniques have been described for identifying particular gene sequences on the basis of their gene products. For example, several techniques are described in the art [see, e.g., International Patent Application No. WO91/07087, published May 30, 1991]. Additionally, known methods exist for the amplification of desired sequences [see, e.g., International Patent Application No. WO91/17271,
10 published November 14, 1991, among others].

However, at present, there exist no established methods for filling the need in the art for methods and reagents which employ fragments of differentially expressed genes of known, unknown (or previously unrecognized) function or consequence to provide diagnostic and therapeutic methods and reagents for diagnosis
15 and treatment of disease or infection, which conditions are characterized by such genes and gene products. It should be appreciated that it is the expression differences that are diagnostic of the altered state (e.g., predisease, disease, pathogenic, progression or infectious). Such genes associated with the altered state are likely to be the targets of drug discovery, whether the genes are the cause or the effect of the
20 condition, identification of such genes provides insight into which gene expression needs to be re-altered in order to reestablished the healthy state.

Summary of the Invention

In one aspect, the invention provides methods for identifying gene(s)
25 which are differentially expressed, for example, in a normal healthy organism and an organism having a disease. The method involves producing and comparing hybridization patterns formed between samples of expressed mRNA or cDNA polynucleotide sequences obtained from either analogous cells, tissues or organs of a healthy organism and a diseased organism and a defined set of
30 oligonucleotide/polynucleotide/polynucleotide sequence probes from either an healthy organism or a diseased organism immobilized on a support. Those defined oligonucleotide/polynucleotide sequences are representative of the total expressed genetic component of the cells, tissues, organs or organism as defined the collection of partial cDNA sequences (ESTs). The differences between the hybridization
35 patterns permit identification of those particular EST or gene-specific oligonucleotide/polynucleotide sequences associated with differential expression, and the identification of the EST permits identification of the clone from which it was

derived and using ordinary skill further cloning and, if desired, sequencing of the full-length cDNA and genomic counterpart, i.e., gene, from which it was obtained.

5 In another aspect, the invention provides methods substantially similar to those described above, but which permit identification of those gene(s) of a pathogen which are expressed in any biological sample of an infected organism based on comparative hybridization of RNA/cDNA samples derived from a healthy versus infected organism, hybridized to an oligonucleotide/polynucleotide set representative of the gene coding complement of the pathogen of interest.

10 In another aspect, the invention provides methods substantially similar to those described above, but which permit identification of those ESTs-specific oligonucleotide/polynucleotide sequences of host gene(s) which represent genes being differentially expressed/ altered in expression by the disease state, or infection and are expressed in any biological sample of an infected organism based on comparative hybridization of RNA/cDNA samples derived from a healthy versus infected
15 organism of interest.

In a further aspect, the methods described above and in detail below, also provide methods for diagnosis of diseases or infections characterized by differentially expressed genes, the expression of which has been altered as a result of infection by the pathogen or disease causing agent in question. All identified
20 differences provide the basis for diagnostic testing be it the altered expression of endogenous genes or the patterned expression of the genes of the infecting organism. Such patterns of altered expression are defined by comparing RNA/cDNA from the two states hybridized against a panel of oligonucleotide/polynucleotides representing the expressed gene component of a cell, tissue, organ or organism as defined by its
25 collection of ESTs.

Yet a further aspect of this invention provides a composition suitable for use in hybridization, which comprises a solid surface on which is immobilized at pre-defined regions thereon a plurality of defined oligonucleotide/polynucleotide sequences for hybridization, each sequence comprising a fragment of an EST isolated
30 from a cDNA or DNA library prepared from at least one selected tissue or cell sample of a healthy (i.e., pre-disease state) animal, at least one analogous sample of an animal having a disease, at least one analogous sample of an animal infected with a pathogen or the pathogen itself, or any combination or multiple combinations thereof.

An additional aspect of the invention provides an isolated gene
35 sequence which is differentially expressed in a normal healthy animal and an animal having a disease, and is identified by the methods above. Similarly, an isolated pathogen gene sequence which is expressed in tissue or cell samples of an infected animal can be identified by the methods above.

Yet another aspect of the invention is that it provides not only a means for a static diagnostic but also provides a means for a carrying out the procedure over time to measure disease progression as well as monitoring the efficacy of disease treatment regimes including an toxicological effects thereof.

5 Another aspect of the invention is an isolated protein produced by expression of the gene sequences identified above. Such proteins are useful in therapeutic compositions or diagnostic compositions, or as targets for drug development.

10 Other aspects and advantages of the present invention are described further in the following detailed description of the preferred embodiments thereof.

Detailed Description of the Invention

15 The present invention meets the unfulfilled needs in the art by providing methods for the identification and use of gene fragments and genes, even those of unknown full length sequence and unknown function, which are differentially expressed in a healthy animal and in an animal having a specific disease or infection by use of ESTs derived from DNA libraries of healthy and/or diseased/infected animals. Employing the methods of this invention permits the resulting identification and isolation of such genes by using their corresponding ESTs
20 and thereby also permits the production of protein products encoded by such genes. The genes themselves and/or protein products, if desired, may be employed in the diagnosis or therapy of the disease or infection with which the genes are associated and in the development of new drugs therefor.

25 It has been appreciated that one or more differentially identified EST or gene-specific oligonucleotide/polynucleotides define a pattern of differentially expressed genes diagnostic of a predisease, disease or infective state. A knowledge of the specific biological function of the EST is not required only that the ESTs identifies a gene or genes whose altered expression is associated reproducibly with the predisease, disease or infectious state. The differences permit the identification of
30 gene products altered in their expression by the disease and represent those products most likely to be targets of therapeutic intervention. Similarly, the product may be of the infecting organism itself and also be an effective target of intervention.

I. Definitions.

35 Several words and phrases used throughout this specification are defined as follows:

As used herein, the term "gene" refers to the genomic nucleotide sequence from which a cDNA sequence is derived, which cDNA produces an EST, as

described below. The term gene classically refers to the genomic sequence, which, upon processing, can produce different cDNAs, e.g., by splicing events. However, for ease of reading, any full-length counterpart cDNA sequence which gives rise to an EST will also be referred to by shorthand herein as a 'gene'.

5 The term "organism" includes without limitation, microbes, plants and animals.

 The term "animal" is used in its broadest sense to include all members of the animal kingdom, including humans. It should be understood, however, that according to this invention the same species of animal which provides the biological
10 sample also is the source of the defined immobilized oligonucleotide/polynucleotides as defined below.

 The term "pathogen" is defined herein as any molecule or organism which is capable of infecting an animal or plant and replicating its nucleic acid sequences in the cells or tissues of that animal or plant. Such a pathogen is generally
15 associated with a disease condition in the infected animal or plant. Such pathogens may include viruses, which replicate intra- or extra-cellularly, or other organisms, such as bacteria, fungi or parasites, which generally infect tissues or the blood. Certain pathogens or microorganisms are known to exist in sequential and distinguishable stages of development, e.g., latent stages, infective stages, and stages
20 which cause symptomatic diseases. In these different stages, the pathogens are anticipated to express differentially certain genes and/or turn on or off host cell gene expression.

 As used herein, the term "disease" or "disease state" refers to any condition which deviates from a normal or standardized healthy state in an organism
25 of the same species in terms of differential expression of the organism's genes. In other words, a disease state can be any illness or disorder be it of genetic or environmental origin, for example, an inherited disorder such as certain breast cancers, or a disorder which is characterized by expression of gene(s) normally in an inactive, 'turned off' state in a healthy animal, or a disorder which is characterized by
30 under-expression or no expression of gene(s) which is normally activated or 'turned on' in a normal healthy animal. Such differential expression of genes may also be detected in a condition caused by infection, inflammation, or allergy, a condition caused by development or aging of the animal, a condition caused by administration of a drug or exposure of the animal to another agent, e.g., nutrition, which affects
35 gene expression. Essentially, the methods described herein can be adapted to detect differential gene expression resulting from any cause, by manipulation of the defined oligonucleotide/polynucleotides and the samples tested as described below. The

concept of disease or disease state also includes its temporal aspects in terms of progression and treatment.

The phrase "differentially expressed" refers to those situations in which a gene transcript is found in differing numbers of copies, or in activated vs inactivated states, in different cell types or tissue types of an organism, having a selected disease as contrasted to the levels of the gene transcript found in the same cells or tissues of a healthy organism. Genes may be differentially expressed in differing states of activation in microorganisms or pathogens in different stages of development. For example, multiple copies of gene transcripts may be found in an organism having a selected disease, while only one, or significantly fewer copies, of the same gene transcript are found in a healthy organism, or vice-versa.

As used herein, the term "solid support" refers to any known substrate which is useful for the immobilization of large numbers of oligonucleotide/polynucleotide sequences by any available method to enable detectable hybridization of the immobilized oligonucleotide/polynucleotide sequences with other polynucleotide sequences in a sample. Among a number of available solid supports, one desirable example is the supports described in International Patent Application No. WO91/07087, published May 30, 1991. Also useful are supports such as but not limited to nitrocellulose, mylein, glass, silica and Pall Biodyne C®. It is also anticipated that improvements yet to be made to conventional solid supports may also be employed in this invention.

The term "surface" means any generally two-dimensional structure on a solid support to which the desired oligonucleotide/polynucleotide sequence is attached or immobilized. A surface may have steps, ridges, kinks, terraces and the like.

As used herein, the term "predefined region" refers to a localized area on a surface of a solid support on which is immobilized one or multiple copies of a particular oligonucleotide/polynucleotide sequence and which enables the identification of the oligonucleotide/polynucleotide at the position, if hybridization of that oligonucleotide/polynucleotide to a sample polynucleotide occurs.

By "immobilized" refers to the attachment of the oligonucleotide/polynucleotide to the solid support. Means of immobilization are known and conventional to those of skill in the art, and may depend on the type of support being used.

By "EST" or "Expressed Sequence Tag" is meant a partial DNA or cDNA sequence of about 150 to 500, more preferably about 300, sequential nucleotides of a longer sequence obtained from a genomic or cDNA library prepared from a selected cell, cell type, tissue or tissue type, organ or organism which longer

sequence corresponds to an mRNA of a gene found in that library. An EST is generally DNA. One or more libraries made from a single tissue type typically provide at least about 3000 different (i.e., unique) ESTs and potentially the full complement of all possible ESTs representing all cDNAs e.g., 50,000-100,000 in an animal such as a human. Further background and information on the construction of ESTs is described in M. D. Adams et al, Science, 252:1651-1656 (1991); and International Application Number PCT/US92/05222 (January 7, 1993).

As used herein, the term "defined oligonucleotide/polynucleotide sequence" refers to a known nucleotide sequence fragment of a selected EST or gene. This term is used interchangeably with the term "fragments of EST". These sequential sequences are generally comprised of between about 15 to about 45 nucleotides and more preferably between about 20 to about 25 nucleotides in length. Thus any single EST of 300 nucleotides in length may provide about 280 different defined oligonucleotide/polynucleotide sequences of 20 nucleotides in length (e.g., 20-mers). The lengths of the defined oligonucleotide/polynucleotides may be readily increased or decreased as desired or needed, depending on the limitations of the solid support on which they may be immobilized or the requirements of the hybridization conditions to be employed. The length is generally guided by the principle that it should be of sufficient length to insure that it is one average only represented once in the population to be examined. Generally, these defined oligonucleotide/polynucleotides are RNA or DNA and are preferably derived from the anti-sense strand of the EST sequence or from a corresponding mRNA sequence to enable their hybridization with samples of RNA or DNA. Modified nucleotides may be incorporated to increase stability and hybridization properties.

By the term "plurality of defined oligonucleotide/polynucleotide sequences" is meant the following. A surface of a solid support may immobilize a large number of "defined oligonucleotide/polynucleotides". For example, depending upon the nature of the surface, it can immobilize from about 300 to upwards of 60,000 defined 20-mer oligonucleotide/polynucleotides. It is anticipated that future improvements to solid surfaces will permit considerably larger such pluralities to be immobilized on a single surface. A "plurality" of sequences refers to the use on any one solid support of multiple different defined oligonucleotide/polynucleotides from a single EST from a selected library, as well as multiple different defined oligonucleotide/polynucleotides from different ESTs from the same library or many libraries from the same or different tissues, and may also include multiple identical copies of defined oligonucleotide/polynucleotides. Ultimately a plurality has at least one oligonucleotide/polynucleotide per expressed gene in the entire organism. For example, from a library producing about 5,000-10,000 ESTs, a single support can

include at least about 1-20 defined oligonucleotide/polynucleotides representing every EST in that library. The composition of defined oligonucleotide/polynucleotides which make up a surface according to this invention may be selected or designed as desired.

5 The term "sample" is employed in the description of this invention in several important ways. As used herein, the term "sample" encompasses any cell or tissue from an organism. Any desired cell or tissue type in any desired state may be selected to form a sample. For example, the sample cell desired may be a human T cell; the desired cell type for use in this invention may be a quiescent T cell or an
10 activated T cell.

 By the phrase "analogous sample" or "analogous cell or tissue" is meant that according to this invention when the ESTs which provide the defined oligonucleotide/polynucleotides are produced from a cDNA library prepared from a single tissue or cell type source sample, e.g., liver tissue of a human, then the samples
15 used to hybridize to those immobilized defined oligonucleotide/polynucleotides are preferably provided by the same type of sample from either a healthy or diseased animal, i.e., liver tissue of a healthy human and liver tissue of a diseased or infected human or from a human suspected of having that disease or infection. Alternatively, if the surface contains defined oligonucleotide/polynucleotides from multiple cells or
20 tissues, then the "samples" which are hybridized thereto can be but are not limited to samples obtained from analogous multiple tissues or cells.

 By the term "detectably hybridizing" means that the sample from the healthy organism or diseased or infected organism is contacted with the defined oligonucleotide/polynucleotides on the surface for sufficient time to permit the
25 formation of patterns of hybridization on the surfaces caused by hybridization between certain polynucleotide sequences in the samples with the certain immobilized defined oligonucleotide/polynucleotides. These patterns are made detectable by the use of available conventional techniques, such as fluorescent labelling of the samples. Preferably hybridization takes place under stringent conditions, e.g., revealing
30 homologies of about 95%. However, if desired, other less stringent conditions may be selected. Techniques and conditions for hybridization at selected stringencies are well known in the art [see, e.g., Sambrook et al, Molecular Cloning. A Laboratory Manual, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY (1989)].

35 II. Compositions of The Invention

 The present invention is based upon the use of ESTs from any desired cell or tissue in known technologies for oligonucleotide/polynucleotide hybridization.

A. *ESTs*

An EST, as defined above, is for an animal, a sequence from a cDNA clone that corresponds to an mRNA. The EST sequences useful in the present invention are isolated preferably from cDNA libraries using a rapid screening and sequencing technique. Custom made cDNA libraries are made using known techniques. See, generally, Sambrook et al, cited above. Briefly, mRNA from a selected cell or tissue is reverse transcribed into complementary DNA (cDNA) using the reverse transcriptase enzyme and made double-stranded using RNase H coupled with DNA polymerase or reverse transcriptase. Restriction enzyme sites are added to the cDNA and it is cloned into a vector. The result is a cDNA library. Alternatively, commercially available cDNA libraries may be used. Libraries of cDNA can also be generated from recombinant expression of genomic DNA using known techniques, including polymerase chain reaction-derived techniques.

ESTs (which can range from about 150 to about 500 nucleotides in length, preferably about 300 nucleotides) can be obtained through sequence analysis from either end of the cDNA insert. Desirably, the DNA libraries used to obtain ESTs use directional cloning methods so that either the 5' end of the cDNA (likely to contain coding sequence) or the 3' end (likely to be a non-coding sequence) can be selectively obtained.

In general, the method for obtaining ESTs comprises applying conventional automated DNA sequencing technology to screen clones, advantageously randomly selected clones, from a cDNA library. The cDNA libraries from the desired tissue can be preprocessed, or edited, by conventional techniques to reduce repeated sequencing of high and intermediate abundance clones and to maximize the chances of finding rare messages from specific cell populations. Preferably, preprocessing includes the use of defined composition prescreening probes, e.g., cDNA corresponding to mitochondria, abundant sequences, ribosomes, actins, myelin basic polypeptides, or any other known high abundance peptide. These prescreening probes used for preprocessing are generally derived from known ESTs. Other useful preprocessing techniques include subtraction hybridization, which preferentially reduces the population of highly represented sequences in the library [e.g., see Fargnoli et al, *Anal. Biochem.* 187:364 (1990)] and normalization, which results in all sequences being represented in approximately equal proportions in the library [Patanjali et al, *Proc. Natl. Acad. Sci. USA*, 88:1943 (1991)]. Additional prescreening/differential screening approaches are known to those skilled in the art.

ESTs can then be generated from partial DNA sequencing of the selected clones. The ESTs useful in the present invention are preferably generated using low redundancy of sequencing, typically a single sequencing reaction. While

single sequencing reactions may have an accuracy as low as 90%, this nevertheless provides sufficient fidelity for identification of the sequence and design of PCR primers.

If desired, the location of an EST in a full length cDNA is determined by analyzing the EST for the presence of coding sequence. A conventional computer program is used to predict the extent and orientation of the coding region of a sequence (using all six reading frames). Based on this information, it is possible to infer the presence of start or stop codons within a sequence and whether the sequence is completely coding or completely non-coding or a combination of the two. If start or stop codons are present, then the EST can cover both part of the 5'-untranslated or 3'-untranslated part of the mRNA (respectively) as well as part of the coding sequence. If no coding sequence is present, it is likely that the EST is derived from the 3' untranslated sequence due to its longer length and the fact that most cDNA library construction methods are biased toward the 3' end of the mRNA. It should be understood that both coding and non-coding regions may provide ESTs equally useful in the described invention.

A number of specific ESTs suitable for use in the present invention are described above Adams et al (*supra*), which may be incorporated by reference herein, to describe non-essential examples of desirable ESTs. Other ESTs exist in the art which may also be useful in this invention, as will ESTs yet to be developed by these known techniques.

B. Preparing the Solid Support of the Invention

Oligonucleotide sequences which are fragments of defined sequence are derived from each EST by conventional means, e.g., conventional chemical synthesis or recombinant techniques. Each defined oligonucleotide/polynucleotide sequence as described above is a fragment, can be, but is not necessarily an anti-sense fragment, of an EST isolated from a DNA library prepared from a selected cell or tissue type from a selected animal. For use in the present invention, it is presently preferred that the defined oligonucleotide/polynucleotide sequences are 20-25mers. As described above, for each EST a number of such 20-25mers may be generated. The lengths may vary as described above as well as the composition. For example oligonucleotide/polynucleotides can be modified based on the Oligo 4.0 or similiar programs to predict hybridization potential or to include modified nucleotides for the reasons given above. It is also appreciated that large DNA segments may be employed including entire ESTs or even full length genes particular when inserted into cloning vectors.

A plurality of these defined oligonucleotide/polynucleotide sequences are then attached to a selected solid support conventionally used for the attachment of nucleotide sequences again by known means. In contrast to other technologies available in the art, this support is designed to contain defined, not random, oligonucleotide/polynucleotide sequences. The EST fragments, or defined oligonucleotide/polynucleotide sequences, immobilized on the solid support can include fragments of one or more ESTs from a library of at least one selected tissue or cell sample of a healthy animal, at least one analogous sample of the animal having a disease, at least one analogous sample of the animal infected with a pathogen, and any combination thereof.

Numerous conventional methods are employed for attaching biological molecules such as oligonucleotide/polynucleotide sequences to surfaces of a variety of solid supports. See, e.g., Affinity Techniques, Enzyme Purification: Part B, Methods in Enzymology, Vol. 34, ed. W.B. Jakoby, M. Wilcheck, Acad. Press, NY (1974); Immobilized Biochemicals and Affinity Chromatography, Advances in Experimental Medicine and Biology, vol. 42, ed. R. Dunlap, Plenum Press, NY (1974); U. S. Patent No. 4,762,881; U. S. Patent No. 4,542,102; European Patent Publication No. 391,608 (October 10, 1990); U. S. Patent No. 4,992,127 (Nov. 21, 1989).

One desirable method for attaching oligonucleotide/polynucleotide sequences derived from ESTs to a solid support is described in International Application No. PCT/US90/06607 (published May 30, 1991). Briefly, this method involves forming predefined regions on a surface of a solid support, where the predefined regions are capable of immobilizing ESTs. The methods make use of binding substances attached to the surface which enable selective activation of the predefined regions. Upon activation, these binding substances become capable of binding and immobilizing oligonucleotide/polynucleotides based on EST or longer gene sequences.

Any of the known solid substrates suitable for binding oligonucleotide/polynucleotides at pre-defined regions on the surface thereof for hybridization and methods for attaching the oligonucleotide/polynucleotides thereto may be employed by one of skill in the art according to this invention. Similarly, known conventional methods for making hybridization of the immobilized oligonucleotide/polynucleotides detectable, e.g., fluorescence, radioactivity, photoactivation, biotinylation, solid state circuitry, and the like may be used in this invention.

Thus, by resorting to known techniques, the invention provides a composition suitable for use in hybridization which consists of a surface of a solid

support on which is immobilized at pre-defined regions on said surface a plurality of defined oligonucleotide/polynucleotide sequences for hybridization. For example, one composition of this invention is a solid support on which are immobilized oligos of EST fragments from a library constructed from a single cell type, e.g., a human stem cell, or a single tissue, e.g., human liver, from a healthy human. Still another composition of this invention is another solid support on which are immobilized oligos of EST fragments from a library constructed from a single cell type or a tissue from a human having a selected disease or predisposition to a selected disease, e.g., liver cancer.

Another embodiment of the compositions of this invention include a single solid support having oligonucleotides of ESTs from both single cell or single tissue libraries from both a healthy and diseased human. Still other embodiments include a single support on which are immobilized oligos of EST fragments from more than one tissue or cell library from a healthy human or a single support on which are immobilized more than one tissue or cell library from both healthy and diseased animals or humans. A preferred composition of this invention is anticipated to be a single support containing oligos of ESTs for all known cells and tissues from a selected organism.

III. The Methods of the Invention

A. Identification of Genes

The present invention employs the compositions described above in methods for identifying genes which are differentially expressed in a normal healthy organism and an organism having a disease or infection. These methods may be employed to detect such genes, regardless of the state of knowledge about the function of the gene. The method of this invention by use of the compositions containing multiple defined EST fragments from a single gene as described above is able to detect levels of expression of genes or in other cases simply the expression or lack thereof, which differ between normal, healthy organisms and organisms having a selected disease, disorder or infection.

One such method employs a first surface of a solid support on which is immobilized at pre-defined regions thereon a plurality of defined oligonucleotide/polynucleotide sequences, described above, of EST or longer gene fragment isolated from a cDNA library prepared from at least one selected tissue or cell sample of a healthy animal (the "healthy test surface") and a second such surface on which is immobilized at pre-defined regions a plurality of defined oligonucleotide/polynucleotide sequences of EST or longer gene fragment isolated from at least one analogous tissue of an animal having a selected disease (the "disease

test surface"). These test surfaces may be standardized for the selected animal or selected cell or tissue sample from that animal (i.e., they are prescreened for polymorphisms in the species population).

Polynucleotide sequences are then isolated from mRNA and/or
5 cDNA from a biological sample from a known healthy animal ("healthy control") and a second sample is similarly prepared from a sample from a known diseased animal ("disease sample"). These two samples are desirably selected from the cell or tissue analogous to that which provided the immobilized oligonucleotide/polynucleotides.

According to the method the healthy control sample is
10 contacted with one set of the healthy test surface and the disease test surface described above for a time sufficient to permit detectable hybridization to occur between the sample and the immobilized defined oligonucleotide/polynucleotides on each surface. The results of this hybridization are a first hybridization pattern formed between the nucleotides of healthy control and the healthy test surface and a second
15 hybridization pattern formed between the nucleotides of healthy control sample and the disease test surface.

In a similar manner, the disease sample is detectably hybridized to another set of healthy test and disease test surfaces, forming a third hybridization pattern between the disease sample and healthy test surface and a fourth hybridization
20 pattern between the disease sample and the disease test surface.

Comparing the four hybridization patterns permits detection of those defined oligonucleotide/polynucleotides which are differentially expressed between the healthy control and the disease sample by the presence of differences in the hybridization patterns at pre-defined regions. The
25 oligonucleotide/polynucleotides on each surface which correspond to the pattern differences may be readily identified with the corresponding EST or longer gene fragment from which the oligonucleotide/polynucleotides are obtained.

In another embodiment of the method of this invention, the same process is employed, with the exception that plurality of defined
30 oligonucleotide/polynucleotide sequences forming the healthy test sample and the disease test sample surfaces are immobilized on a single solid support. For example, each fragment of an EST or longer gene fragment on the surface is isolated from at least two cDNA libraries prepared from a selected cell or tissue sample of a healthy animal and an analogous selected cell or tissue sample of an animal having a disease.

35 According to this embodiment, the healthy control sample is detectably hybridized to a copy of this single solid surface, forming one hybridization pattern with oligonucleotide/polynucleotides associated with both the healthy and diseased animal. Similarly, the disease sample is detectably hybridized to a second

copy of this single solid surface, forming one hybridization pattern with oligonucleotide/polynucleotides associated with both the healthy and diseased animal.

Comparing the two hybridization patterns permits detection of those defined oligonucleotide/polynucleotides which are differentially expressed between the healthy control and the disease sample by the presence of differences in the hybridization patterns at pre-defined regions. The oligonucleotide/polynucleotides on each surface which correspond to the pattern differences may be readily identified with the corresponding EST or longer gene fragment from which the oligonucleotide/polynucleotides are obtained.

The identification of one or more ESTs as the source of the defined oligonucleotide/polynucleotide which produced a "difference" in hybridization patterns according to these methods permits ready identification of the gene from which those ESTs were derived. Because oligonucleotides are of sufficient length that they will hybridize under stringent conditions only with a RNA/cDNA for that gene to which they correspond, the oligo can be used to identify the EST and in turn the clone from which it was derived and by subsequent cloning, obtain the sequence of the full-length cDNA and its genomic counterparts, i.e., the gene, from which it was obtained.

In other words, the ESTs identified by the method of this invention can be employed to determine the complete sequence of the mRNA, in the form of transcribed cDNA, by using the EST as a probe to identify a cDNA clone corresponding to a full-length transcript, followed by sequencing of that clone. The EST or the full length cDNA clone can also be used as a probe to identify a genomic clone or clones that contain the complete gene including regulatory and promoter regions, exons, and introns.

It should be appreciated that one does not have to be restricted in using ESTs from a particular tissue from which probe RNA or cDNA is obtained, rather any or all ESTs (known or unknown) may be placed on the support. Hybridization will be used as a form of diagnostic patterns or to identify which particular EST is detected. For example, all known ESTs from an organism are used to produce a "master" solid support to which control sample and disease samples are alternately hybridized. One then detects a pattern of hybridization associated with the particular disease state which then forms the basis of a diagnostic test or the isolation of disease specific ESTs from which the intact gene may be cloned and sequenced leading ultimately to a defined therapeutic target.

Methods for obtaining complete gene sequences from ESTs are well-known to those of skill in the art. See, generally, Sambrook et al, cited above. Briefly, one suitable method involves purifying the DNA from the clone that was

sequenced to give the EST and labeling the isolated insert DNA. Suitable labeling systems are well known to those of skill in the art [see, eg. Basic Methods in Molecular Biology, L. G. Davis et al, ed., Elsevier Press, NY (1986)]. The labeled EST insert is then used as a probe to screen a lambda phage cDNA library or a plasmid cDNA library, identifying colonies containing clones related to the probe cDNA which can be purified by known methods. The ends of the newly purified clones are then sequenced to identify full length sequences and complete sequencing of full length clones is performed by enzymatic digestion or primer walking. A similar screening and clone selection approach can be applied to clones from a genomic DNA library.

Additionally, an EST or gene identified by this method as associated with inherited disorders can be used to determine at what stage during embryonic development the selected gene from which it is derived is developed by screening embryonic DNA libraries from various stages of development, e.g. 2-cell, 8-cell, etc., for the selected gene. As has been mentioned above, the invention may be applied in additional temporal modes for monitoring the progression of a disease state, the efficacy of a particular treatment modality or the aging process of an individual.

Thus, the methods of this invention permit the identification, isolation and sequencing of a gene which is differentially expressed in a selected disease/infection. As described in more detail below, the identified gene may then be employed to obtain any protein encoded thereby, or may be employed as a target for diagnostic methods or therapeutic approaches to the treatment of the disease, including, e.g., drug development.

The same methods as described above for the identification of genes, including genes of unknown function, which are differentially expressed in a disease state, may also be employed to identify other genes of interest. For example, another embodiment of this invention includes a method for identifying a gene of a pathogen which is expressed in a biological sample of an animal infected with that pathogen or the gene of the host which is altered in its expression as a result of the infection.

One such method employs a healthy test surface as described above, employing defined oligonucleotide/polynucleotides from a sample of a healthy, uninfected animal. The second such surface has immobilized at pre-defined regions thereon a plurality of defined oligonucleotide/polynucleotide sequences of ESTs isolated from at least one analogous tissue or cell sample of an infected animal (the "infection test surface"). Polynucleotide sequences are isolated from a biological sample from a healthy animal ("healthy control") and a second sample is similarly

prepared from an animal infected with the selected pathogen ("infection sample"). These two samples are desirably selected from the cell or tissue analogous to that which provided the immobilized oligonucleotide/polynucleotides. It would also be possible to provide samples from the nucleic acid of the pathogen itself.

5 According to the method the healthy control sample is contacted with one set of the healthy test surface and the infection test surface described above for a time sufficient to permit detectable hybridization to occur between the sample and the immobilized defined oligonucleotide/polynucleotides on each surface. The results of this hybridization are a first hybridization pattern formed
10 between the nucleotides of healthy control and the healthy test surface and a second hybridization pattern formed between the nucleotides of healthy control sample and the infection test surface.

 In a similar manner, the infection sample is detectably hybridized to another set of healthy test and infection test surfaces, forming a third
15 hybridization pattern between the infection sample and healthy test surface and a fourth hybridization pattern between the infection sample and the infection test surface.

 Comparing the four hybridization patterns permits detection of those defined oligonucleotide/polynucleotides which are differentially expressed
20 between the healthy animal and the animal infected with the pathogen by the presence of differences in the hybridization patterns at pre-defined regions. As mentioned differential expression is not required and simple qualitative analysis is possible by reference to gene expression which is simply present or absent.

 A second embodiment of this method parallels the second
25 embodiment of the method as applied to disease above, i.e., the same process is employed, with the exception that plurality of defined oligonucleotide/polynucleotide sequences forming the healthy test sample surface and the infection test sample surface are immobilized on a single solid support. The resulting first hybridization pattern (healthy control sample with healthy/infection test sample) and second
30 hybridization pattern (infection sample with healthy/infection test sample) permits detection of those defined oligonucleotide/polynucleotides which are differentially expressed between the healthy control and the infection sample by the presence of differences in the hybridization patterns at pre-defined regions. The oligonucleotide/polynucleotides on each surface which correspond to the pattern
35 differences may be readily identified with the corresponding ESTs from which the oligonucleotide/polynucleotides are obtained.

 As described above for the methods for identifying differential gene expression between diseased and healthy animals, the

oligonucleotide/polynucleotides on each surface which correspond to the pattern differences may be readily identified with the corresponding ESTs from which the oligonucleotide/polynucleotide sequences are obtained and the genes expressed by the pathogen identified for similar purposes. Other embodiments of these methods may be developed with resort to the teaching herein, by altering the samples which provide the defined oligonucleotide/polynucleotides. For example, an EST, identified with a differentially expressed gene by the method of this invention is also useful in detecting genes expressed in the various stages of an pathogen's development, particularly the infective stage and following the cours of drug treatment and emergence of resistant variants. For example, employing the techniques described above, the EST can be used for detecting a gene in various stages of the parasitic *Plasmodium* species life cycle, which include blood stages, liver stages, and gametocyte stages.

B. Diagnostic Methods

In addition to use of the methods and compositions of this invention for identifying differentially expressed genes, another embodiment of this invention provides diagnostic methods for diagnosing a selected disease state, or a selected state resulting from aging, exposure to drugs or infection in an animal. According to this aspect of the invention, a first surface, described as the healthy test surface above, and a second surface, described as the disease test surface or infection test surface, are prepared depending on the disease or infection to be diagnosed. The same processes of detectable hybridization to a first and second set of these surfaces with the healthy control sample and disease/infection sample are followed to provide the four above-described hybridization patterns, i.e., healthy control sample with healthy test surface; healthy control sample with disease/infection test surface; disease/infection sample with healthy test surface; and disease/infection sample with disease/infection test surface.

The diagnosis of disease or infection is provided by comparing the four hybridization patterns. Substantial differences between the first and third hybridization patterns, respectively, and the second and fourth hybridization patterns, respectively, indicate the presence of the selected disease or infection in said animal. Substantial similarities in the first and third hybridization patterns and second and fourth hybridization patterns indicates the absence of disease or infection.

A similar embodiment utilizes the single surface bearing both the healthy test surface defined oligonucleotide/polynucleotides and the disease/infection test surface defined oligonucleotide/polynucleotides as described above. Parallel process steps as described above for detection of genes differentially expressed in disease and infected states are followed, resulting in a first hybridization

pattern (healthy control sample with single healthy and disease/infection test sample) and a second hybridization pattern (disease/infection sample with another copy of the single healthy and disease/infection test sample).

Diagnosis is accomplished by comparing the two hybridization patterns, wherein substantial differences between the first and second hybridization patterns indicate the presence of the selected disease or infection in the animal being tested. Substantially similar first and second hybridization patterns indicate the absence of disease or infection. This like many of the foregoing embodiments may use known or unknown ESTs derived from many libraries.

10 C. *Other Methods of the Invention*

As is obvious to one of skill in the art upon reading this disclosure, the compositions and methods of this invention may also be used for other similar purposes. For example, the general methods and compositions may be adapted easily by manipulation of the samples selected to provide the standardized defined oligonucleotide/polynucleotides, and selection of the samples selected for hybridization thereto. One such modification is the use of this invention to identify cell markers of any type, e.g., markers of cancer cells, stem cell markers, and the like. Another modification involves the use of the method and compositions to generate hybridization patterns useful for forensic identification or an 'expression fingerprint' of genes for identification of one member of a species from another. Similarly, the methods of this invention may be adapted for use in tissue matching for transplantation purposes as well as for molecular histology, i.e., to enable diagnosis of disease or disorders in pathology tissue samples such as biopsies. Still another use of this method is in monitoring the effects of development and aging upon the gene expression in a selected animal, by preparing surfaces bearing oligonucleotide/polynucleotides prepared from samples of standardized younger members of the species being tested. Additionally the patient can serve as an internal control by virtue of having the method applied to blood samples every 5-10 years during his lifetime.

Still another intriguing use of this method is in the area of monitoring the effects of drugs on gene expression, both in laboratories and during clinical trials with animal, especially humans. Because the method can be readily adapted by altering the above parameters, it can essentially be employed to identify differentially expressed genes of any organism, at any stage of development, and under the influence of any factor which can affect gene expression.

IV. *The Genes and Proteins Identified*

Application of the compositions and methods of this invention as above described also provide other compositions, such as any isolated gene sequence which is differentially expressed between a normal healthy animal and an animal
5 having a disease or infection. Another embodiment of this invention is any isolated pathogen gene sequence which is expressed in tissue or cell samples of an infected animal. Similarly an embodiment of this invention is any gene sequence identified by the methods described herein.

These gene sequences may be employed in conventional methods to
10 produce isolated proteins encoded thereby. To produce a protein of this invention, the DNA sequences of a desired gene identified by the use of the methods of this invention or portions thereof are inserted into a suitable expression system. Desirably, a recombinant molecule or vector is constructed in which the polynucleotide sequence encoding the protein is operably linked to a heterologous
15 expression control sequence permitting expression of the human protein. Numerous types of appropriate expression vectors and host cell systems are known in the art for mammalian (including human) expression, insect, e.g., baculovirus expression, yeast, fungal, and bacterial expression, by standard molecular biology techniques.

The transfection of these vectors into appropriate host cells, whether
20 mammalian, bacterial, fungal, or insect, or into appropriate viruses, can result in expression of the selected proteins. Suitable host cells or cell lines for transfection, and viruses, as well as methods for the construction and transfection of such host cells and viruses are well-known. Suitable methods for transfection, culture, amplification, screening, and product production and purification are also known in the art.

25 The genes and proteins identified by this invention can be employed, if desired in diagnostic compositions useful for the diagnosis of a disease or infection using conventional diagnostic assays. For example, a diagnostic reagent can be developed which detectably targets a gene sequence or protein of this invention in a biological sample of an animal. Such a reagent may be a complementary nucleotide
30 sequence, an antibody (monoclonal, recombinant or polyclonal), or a chemically derived agonist or antagonist. Alternatively, the proteins and polynucleotide sequences of this invention, fragments of same, or complementary sequences thereto, may themselves be useful as diagnostic reagents for diagnosing disease states with which the ESTs of the invention are associated. These reagents may optionally be
35 labelled using diagnostic labels, such as radioactive labels, colorimetric enzyme label systems and the like conventionally used in diagnostic or therapeutic methods, e.g., Northern and Western blotting, antigen-antibody binding and the like. The selection of the appropriate assay format and label system is within the skill of the art and may

readily be chosen without requiring additional explanation by resort to the wealth of art in the diagnostic area.

Additionally, genes and proteins identified according to this invention may be used therapeutically. For example, the EST-containing gene sequences may be useful in gene therapy, to provide a gene sequence which in a disease is not properly or sufficiently expressed. In such a method, a selected gene sequence of this invention is introduced into a suitable vector or other delivery system for delivery to a cell containing a defect in the selected gene. Suitable delivery systems are well known to those of skill in the art and enable the desired EST or gene to be incorporated into the target cell and to be translated by the cell. The EST or gene sequence may be introduced to mutate the existing gene by recombination or provide an active copy thereof in addition to the inactive gene to replace its function.

Alternatively, a protein encoded by an EST or gene of the invention may be useful as a therapeutic reagent for delivery of a biologically active protein, particularly when the disease state is associated with a deficiency of this protein. Such a protein may be incorporated into an appropriate therapeutic formulation, alone or in combination with other active ingredients. Methods of formulating such therapeutic compositions, as well as suitable pharmaceutical carriers, and the like, are well known to those of skill in the art. Still an additional method of delivering the missing protein encoded by an EST, or the gene from which a selected EST was derived, involves expressing it directly *in vivo*. Systems for such *in vivo* expression are well known in the art.

Yet another use of the ESTs, genes identified according to the methods of this invention, or the proteins encoded thereby is a target for the screening and development of natural or synthetic chemical compounds which have utility as therapeutic drugs for the treatment of disease states associated with the identified genes and ESTs derived therefrom. As one example, a compound capable of binding to such a protein encoded by such a gene and either preventing or enhancing its biological activity may be a useful drug component for the treatment or prevention of such disease states.

Conventional assays and techniques may be used for the screening and development of such drugs. As one example, a method for identifying compounds which specifically bind to or inhibit or activate proteins encoded by these gene sequences can include simply the steps of contacting a selected protein or gene product, with a test compound to permit binding of the test compound to the protein; and determining the amount of test compound, if any, which is bound to the protein. Such a method may involve the incubation of the test compound and the protein immobilized on a solid support. Still other conventional methods of drug screening

can involve employing a suitable computer program to determine compounds having similar or complementary chemical structures to that of the gene product or portions thereof and screening those compounds either for competitive binding to the protein to detect enhanced or decreased activity in the presence of the selected compound.

5 Thus, through use of such methods, the present invention is anticipated to provide compounds capable of interacting with these genes, ESTs, or encoded proteins, or fragments thereof, and either enhancing or decreasing the biological activity, as desired. Such compounds are believed to be encompassed by this invention.

10 Numerous modifications and variations of the present invention are included in the above-identified specification and are expected to be obvious to one of skill in the art. Such modifications and alterations to the compositions and processes of the present invention are believed to be encompassed in the scope of the claims appended hereto.

15

WHAT IS CLAIMED IS:

1. A method for identifying genes which are differentially expressed in two different pre-determined states of an organism comprising:
 - 5 a. providing a first surface on which is immobilized at pre-defined regions on said surface a plurality of defined oligonucleotide/polynucleotide sequences, each sequence selected from the group consisting of a fragment of an EST, an entire EST a fragment of a gene or an entire gene, isolated from a DNA library prepared from at least one selected cell, tissue, organ or organism sample in a first
10 state and present in excess relative to the polynucleotide to be hybridized;
 - b. providing a second surface on which is immobilized at pre-defined regions on said surface a plurality of defined oligonucleotide/polynucleotide sequences, each sequence selected from the group consisting of a fragment of an EST, an entire EST a fragment of a gene or an entire gene, isolated from a DNA library
15 prepared from at least one selected cell, tissue, organ or organism sample in a second state and present in excess relative to the polynucleotide to be hybridized;
 - c. detectably hybridizing to a set of said first and second surfaces polynucleotide sequences isolated from a sample from a said organism in said first state, said sample selected from sources analogous to the sources of step (a), said
20 hybridization sufficient to form a first and second hybridization pattern on each said first and second surface,
 - d. detectably hybridizing to a set of said first and second surfaces polynucleotide sequences isolated from a sample from said organism in said second state, said sample selected from sources analogous to the sources of step (c), said
25 hybridization sufficient to form a third and fourth hybridization pattern on each said first and second surface,
 - e. comparing at least two of the four hybridization patterns, wherein genes differentially expressed in said first and second states are identified by the presence of differences in the hybridization patterns at pre-defined regions;
 - 30 f. identifying the oligonucleotide/polynucleotides on each surface which correspond to said pattern differences and the corresponding ESTs or larger gene fragment from which the oligonucleotide/polynucleotides were obtained, whereby identification of the EST or larger gene fragment permits identification of the gene from which the ESTs or larger gene fragment were derived.

35

2. The method according to Claim 1 wherein said first and second states are respectively healthy and disease; pathogen uninfected and pathogen infected; a first progression state and a second progression of a disease or infection; a first treatment state and a second treatment state of a disease or infection; or a first developmental and a second developmental state.

3. The method according to Claim 1 wherein said organism is a plant or an animal.

4. The method according to Claim 3 wherein said animal is a human.

5. A method for identifying genes which are differentially expressed in a normal healthy animal and an animal having a disease comprising:

a. providing a first surface on which is immobilized at pre-defined regions on said surface a plurality of defined oligonucleotide/polynucleotide sequences, each sequence each sequence selected from the group consisting of a fragment of an EST, an entire EST a fragment of a gene or an entire gene, isolated from a DNA library prepared from at least one selected cell, tissue, organ or organism sample in a healthy animal and present in excess relative to the polynucleotide to be hybridized;

b. providing a second surface on which is immobilized at pre-defined regions of said surface a plurality of defined oligonucleotide/polynucleotide sequences, each sequence each sequence selected from the group consisting of a fragment of an EST, an entire EST a fragment of a gene or an entire gene, isolated from a DNA library prepared from at least one selected cell, tissue, organ or organism sample from an animal having said disease and present in excess relative to the polynucleotide to be hybridized;

c. detectably hybridizing to a set of said first and second surfaces polynucleotide sequences isolated from a sample from a healthy animal, said sample selected from sources analogous to the sources of step (a), said hybridization sufficient to form a first and second hybridization pattern on each said first and second surface, said sample selected from a cell or tissue sample analogous to the sample of step (a), said hybridization sufficient to form a first and second hybridization pattern on each said first and second surface;

- d. detectably hybridizing to a set of said first and second surfaces polynucleotide sequences isolated from a sample from an animal having said disease, said sample selected from a cell or tissue sample analogous to the sample of step (c), said hybridization sufficient to form a third and fourth hybridization pattern on each
5 said first and second surface,
- e. comparing at least two of the four hybridization patterns, wherein genes differentially expressed in said first and second states are identified by the presence of differences in the hybridization patterns at pre-defined regions;
- f. identifying the oligonucleotide/polynucleotides on each surface
10 which correspond to said pattern differences and the corresponding ESTs or larger gene fragment from which the oligonucleotide/polynucleotides were obtained, whereby identification of the EST or larger gene fragment permits identification of the gene from which the ESTs or larger gene fragment were derived.
- 15 6. A method for identifying genes which are differentially expressed in a normal healthy animal and an animal having a disease comprising:
- a. providing a surface on which is immobilized at pre-defined regions on said surface a plurality of defined oligonucleotide/polynucleotide sequences, each sequence selected from the group consisting of a fragment of an EST,
20 an entire EST a fragment of a gene or an entire gene isolated from a DNA library prepared from the group selected from at least one selected cell, tissue, organ or organism sample in of a healthy animal and an analogous selected sample of an animal having said disease and both present in excess relative to the polynucleotide to be hybridized;
- 25 b. detectably hybridizing to a first copy of said surface polynucleotide sequences isolated from a healthy animal, said sample selected from a cell or tissue sample analogous to the sample of step (a), said hybridization sufficient to form a first hybridization pattern on said surface;
- c. detectably hybridizing to a second copy of said surface
30 polynucleotide sequences isolated from an animal having said disease, said sample selected from a cell or tissue sample analogous to the sample of step (a), said hybridization sufficient to form a second hybridization pattern on said surface;
- d. comparing the two hybridization patterns, wherein genes differentially expressed in a disease state are identified by the presence of differences
35 in the hybridization patterns at pre-defined regions;

e. identifying the oligonucleotide/polynucleotides on each surface which correspond to said pattern differences and the corresponding ESTs from which the oligonucleotide/polynucleotides are obtained, whereby identification of the EST permits identification of the gene from which the ESTs were derived.

5

7. A method for identifying a gene of a pathogen which is expressed in a biological sample of an animal infected with said pathogen comprising:

a. providing a first surface on which is immobilized at pre-defined regions on said surface a plurality of defined oligonucleotide/polynucleotide sequences, each sequence selected from the group consisting of a fragment of an EST,
10 an entire EST a fragment of a gene or an entire gene isolated from a DNA library prepared from at least one selected cell, tissue, organ or organism sample of a healthy, uninfected animal and present in excess relative to the polynucleotide to be hybridized;

15

b. providing a second surface on which is immobilized at pre-defined regions of said surface a plurality of defined oligonucleotide/polynucleotide sequences, each sequence selected from the group consisting of a fragment of an EST, an entire EST a fragment of a gene or an entire gene isolated from at least one selected cell, tissue, organ or organism sample of an infected animal;

20

c. detectably hybridizing to a set of said first and second surfaces polynucleotide sequences isolated from a sample from a healthy animal, said sample selected from a cell or tissue sample analogous to the sample of step (a), said hybridization sufficient to form first and second hybridization patterns on each said first and second surface,

25

d. detectably hybridizing to a set of said first and second surfaces polynucleotide sequences isolated from a sample from an infected animal, said sample selected from a cell or tissue sample analogous to the sample of step (a), said hybridization sufficient to form third and fourth hybridization patterns on each said first and second surface,

30

e. comparing the four hybridization patterns, wherein genes of said pathogen which are expressed in an infected animal are identified by the presence of differences in the hybridization patterns at pre-defined regions;

f. identifying the oligonucleotide/polynucleotides on each surface which correspond to said pattern differences and the corresponding ESTs from which
35 the oligonucleotide/polynucleotides are obtained, whereby identification of the EST permits identification of the gene from which the ESTs were derived.

8. A method for identifying a gene of a pathogen which is expressed in a biological sample of an animal infected with said pathogen comprising:

- a. providing a surface on which is immobilized at pre-defined regions on said surface a plurality of defined oligonucleotide/polynucleotide sequences, each sequence selected from the group consisting of a fragment of an EST, an entire EST a fragment of a gene or an entire gene isolated from a DNA library prepared from the group selected from at least one selected cell, tissue, organ or organism sample in of a healthy animal and an analogous selected sample of an animal having said disease and both present in excess relative to the polynucleotide to be hybridized
- b. detectably hybridizing to a first copy of said surface polynucleotide sequences isolated from a sample from a healthy animal, said sample selected from a cell or tissue sample analogous to the sample of step (a), said hybridization sufficient to form a first hybridization pattern on said surface;
- c. detectably hybridizing to a second copy of said surface polynucleotide sequences isolated from a sample from an infected animal, said sample selected from a cell or tissue sample analogous to the sample of step (a), said hybridization sufficient to form a second hybridization pattern on said surface;
- d. comparing the two hybridization patterns, wherein genes of said pathogen which are expressed in an infected animal are identified by the presence of differences in the hybridization patterns at pre-defined regions;
- e. identifying the oligonucleotide/polynucleotides on each surface which correspond to said pattern differences and the corresponding ESTs from which the oligonucleotide/polynucleotides are obtained, whereby identification of the EST permits identification of the gene from which the ESTs were derived.

9. A composition suitable for use in hybridization comprising a solid surface on which is immobilized at pre-defined regions on said surface a plurality of defined oligonucleotide/polynucleotide sequences for hybridization, each sequence selected from the group consisting of a fragment of an EST, an entire EST a fragment of a gene or an entire gene isolated from a DNA library prepared from the group selected from at least one selected cell, tissue, organ or organism sample of a healthy animal, at least one analogous sample of said animal having a disease, at least one analogous sample of said animal infected with a microbial pathogen, and any combination thereof.

10. An isolated gene sequence which is differentially expressed in a normal healthy animal and an animal having a disease, identified by the method of claim 1.

5 11. An isolated pathogen gene sequence which is expressed in tissue or cell samples of an infected animal identified by the method of claim 7.

12. A diagnostic composition useful for the diagnosis of a disease comprising a reagent capable of detectably targeting a gene sequence of claim 10 in a biological sample of an animal.

13. A diagnostic composition useful for the diagnosis of infection by a pathogen comprising a reagent capable of detectably targeting a gene sequence of claim 11 in a biological sample of an animal.

15 14. An isolated protein produced by expression of a gene sequence of claim 10.

15. An isolated pathogen protein produced by expression of a gene sequence of claim 11.

16. A therapeutic composition comprising a protein or fragment thereof selected from the group consisting of a protein of claim 10 and a protein of claim 15.

25 17. A method for diagnosing a selected disease or infection in an animal comprising:

a. providing a first surface on which is immobilized at pre-defined regions on said surface a plurality of defined oligonucleotide/polynucleotide sequences, each sequence selected from the group consisting of a fragment of an EST, an entire EST a fragment of a gene or an entire gene, isolated from a DNA library prepared from at least one selected cell, tissue, organ or organism sample of a healthy animal and present in excess relative to the polynucleotide to be hybridized;

b. providing a second surface on which is immobilized at pre-defined regions of said surface a plurality of defined oligonucleotide/polynucleotide sequences, each sequence comprising a fragment of an EST isolated from at least one said tissue of an animal having said disease;

c. detectably hybridizing to a set of said first and second surfaces polynucleotide sequences isolated from a DNA library prepared from a sample from a healthy animal, said sample selected from a cell or tissue sample analogous to the sample of step (a), said hybridization sufficient to form a first and second
5 hybridization pattern on each said first and second surface;

d. detectably hybridizing to a set of said first and second surfaces polynucleotide sequences isolated from a DNA library prepared from a sample from an animal having said disease, said sample selected from a cell or tissue sample analogous to the sample of step (c), said hybridization sufficient to form a third and
10 fourth hybridization pattern on each said first and second surface;

e. comparing the four hybridization patterns, wherein substantial differences between the first and third hybridization patterns and the second and fourth hybridization patterns indicates the presence of said selected disease or infection in said animal, and substantial similarities in said first and third
15 hybridization patterns and second and fourth hybridization patterns indicates the absence of disease or infection.

18. A method for diagnosing a selected disease or infection in an animal comprising:

a. providing a surface on which is immobilized at pre-defined regions on said surface a plurality of defined oligonucleotide/polynucleotide sequences, each sequence comprising a fragment of an EST isolated from a DNA library prepared from the group consisting of a selected cell or tissue sample of a healthy animal and an analogous selected cell or tissue sample of an animal having
20 said disease;

b. detectably hybridizing to a first copy of said surface polynucleotide sequences isolated from a sample from a healthy animal, said sample selected from a cell or tissue sample analogous to the sample of step (a), said hybridization sufficient to form a first hybridization pattern on said surface;

c. detectably hybridizing to a second copy of said surface polynucleotide sequences isolated from a DNA library prepared from a sample from an animal having said disease, said sample selected from a cell or tissue sample analogous to the sample of step (a), said hybridization sufficient to form a second hybridization pattern on said surface;

d. comparing the two hybridization patterns, wherein substantial differences between the first and second hybridization patterns indicates the presence of said selected disease or infection in said animal, and substantial similarities in said first and second hybridization patterns indicates the absence of disease or infection.
35

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US95/01863

A. CLASSIFICATION OF SUBJECT MATTER

IPC(6) : C12Q 1/68

US CL : 435/6

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 435/6

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

APS, CAS, BIOSIS

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	ANALYTICAL BIOCHEMISTRY, VOLUME 187, ISSUED 1990, FARGNOLI ET AL, "LOW-RATIO HYBRIDIZATION SUBTRACTION", PAGES 364-373, SEE ENTIRE DOCUMENT.	1-18
Y	PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES USA, VOLUME 88, ISSUED MARCH 1991, PATANJALI ET AL, "CONSTRUCTION OF A UNIFORM-ABUNDANCE (NORMALIZED) CDNA LIBRARY", PAGES 1943-1947, SEE ENTIRE DOCUMENT.	1-18
Y	SCIENCE, VOLUME 245, ISSUED 29 SEPTEMBER 1989, OLSON ET AL. "A COMMON LANGUAGE FOR PHYSICAL MAPPING OF THE HUMAN GENOME", PAGES 1434-1435, SEE ENTIRE DOCUMENT.	1-18

☒ Further documents are listed in the continuation of Box C.
 ☐ See patent family annex.

* Special categories of cited documents:	*T	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
A document defining the general state of the art which is not considered to be of particular relevance	*X*	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
E earlier document published on or after the international filing date	*Y*	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
L document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	*Z*	document member of the same patent family
O document referring to an oral disclosure, use, exhibition or other means		
P document published prior to the international filing date but later than the priority date claimed		

Date of the actual completion of the international search

03 APRIL 1995

Date of mailing of the international search report

17 MAY 1995

 Name and mailing address of the ISA/US
 Commissioner of Patents and Trademarks
 Box PCT
 Washington, D.C. 20231

Facsimile No. (703) 305-3230

Authorized officer

EGGERTON CAMPBELL

Telephone No. (703) 308-0196

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US95/01863

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	SCIENCE, VOLUME 252, ISSUED 21 JUNE 1991, ADAMS ET AL, "COMPLEMENTARY DNA SEQUENCING: EXPRESSED SEQUENCE TAGS AND HUMAN GENOME PROJECT", PAGES 1651-1656, SEE ENTIRE DOCUMENT.	1-18



INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification ⁶ : C12Q 1/68, G06F 15/00	A1	(11) International Publication Number: WO 95/20681 (43) International Publication Date: 3 August 1995 (03.08.95)
(21) International Application Number: PCT/US95/01160 (22) International Filing Date: 27 January 1995 (27.01.95) (30) Priority Data: 08/187,530 27 January 1994 (27.01.94) US 08/282,955 29 July 1994 (29.07.94) US (71) Applicant: INCYTE PHARMACEUTICALS, INC. [US/US]; 3330 Hillview Avenue, Palo Alto, CA 94304 (US). (72) Inventors: SEILHAMER, Jeffrey, J.; 12555 La Cresta, Los Altos Hills, CA 94022 (US). SCOTT, Randal, W.; 13140 Sun-Mor, Mountain View, CA 94040 (US). (74) Agents: CAGE, Kenneth, L. et al.; Willian Brinks Hofer Gilson & Lione, 2000 K Street, N.W., Suite 200, Washington, DC 20006-1809 (US).		(81) Designated States: AM, AU, BB, BG, BR, BY, CA, CN, CZ, EE, FI, GE, HU, JP, KG, KP, KR, KZ, LK, LR, LT, LV, MD, MG, MN, MX, NO, NZ, PL, RO, RU, SI, SK, TJ, TT, UA, UZ, VN, European patent (AT, BE, CH, DE, DK, ES, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, ML, MR, NE, SN, TD, TG), ARIPO patent (KE, MW, SD, SZ). Published <i>With international search report.</i>
(54) Title: COMPARATIVE GENE TRANSCRIPT ANALYSIS (57) Abstract A method and system for quantifying the relative abundance of gene transcripts in a biological specimen. One embodiment of the method generates high-throughput sequence-specific analysis of multiple RNAs or their corresponding cDNAs (gene transcript imaging analysis). Another embodiment of the method produces a gene transcript imaging analysis by the use of high-throughput cDNA sequence analysis. In addition, the gene transcript imaging can be used to detect or diagnose a particular biological state, disease, or condition which is correlated to the relative abundance of gene transcripts in a given cell or population of cells. The invention provides a method for comparing the gene transcript image analysis from two or more different biological specimens in order to distinguish between the two specimens and identify one or more genes which are differentially expressed between the two specimens.		

344

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AT	Austria	GB	United Kingdom	MR	Mauritania
AU	Australia	GE	Georgia	MW	Malawi
BB	Barbados	GN	Guinea	NE	Niger
BE	Belgium	GR	Greece	NL	Netherlands
BF	Burkina Faso	HU	Hungary	NO	Norway
BG	Bulgaria	IE	Ireland	NZ	New Zealand
BJ	Benin	IT	Italy	PL	Poland
BR	Brazil	JP	Japan	PT	Portugal
BY	Belarus	KE	Kenya	RO	Romania
CA	Canada	KG	Kyrgyzstan	RU	Russian Federation
CF	Central African Republic	KP	Democratic People's Republic of Korea	SD	Sudan
CG	Congo	KR	Republic of Korea	SE	Sweden
CH	Switzerland	KZ	Kazakhstan	SI	Slovenia
CI	Côte d'Ivoire	LI	Liechtenstein	SK	Slovakia
CM	Cameroon	LK	Sri Lanka	SN	Senegal
CN	China	LU	Luxembourg	TD	Chad
CS	Czechoslovakia	LV	Latvia	TG	Togo
CZ	Czech Republic	MC	Monaco	TJ	Tajikistan
DE	Germany	MD	Republic of Moldova	TT	Trinidad and Tobago
DK	Denmark	MG	Madagascar	UA	Ukraine
ES	Spain	ML	Mali	US	United States of America
FI	Finland	MN	Mongolia	UZ	Uzbekistan
FR	France			VN	Viet Nam
GA	Gabon				

COMPARATIVE GENE TRANSCRIPT ANALYSIS

1. FIELD OF INVENTION

The present invention is in the field of molecular biology and computer science; more particularly, the present invention describes methods of analyzing gene transcripts and diagnosing the genetic expression of cells and tissue.

2. BACKGROUND OF THE INVENTION

Until very recently, the history of molecular biology has been written one gene at a time. Scientists have observed the cell's physical changes, isolated mixtures from the cell or its milieu, purified proteins, sequenced proteins and therefrom constructed probes to look for the corresponding gene.

Recently, different nations have set up massive projects to sequence the billions of bases in the human genome. These projects typically begin with dividing the genome into large portions of chromosomes and then determining the sequences of these pieces, which are then analyzed for identity with known proteins or portions thereof, known as motifs. Unfortunately, the majority of genomic DNA does not encode proteins and though it is postulated to have some effect on the cell's ability to make protein, its relevance to medical applications is not understood at this time.

A third methodology involves sequencing only the transcripts encoding the cellular machinery actively involved in making protein, namely the mRNA. The advantage is that the cell has already edited out all the non-coding DNA, and it is relatively easy to identify the protein-coding portion of the RNA. The utility of this approach was not immediately obvious to genomic researchers. In fact, when cDNA sequencing was initially proposed, the method was roundly denounced by those committed to genomic sequencing. For example, the head of the U.S. Human Genome project discounted CDNA sequencing as not valuable and refused to approve funding of projects.

In this disclosure, we teach methods for analyzing DNA, including cDNA libraries. Based on our analyses and

research, we see each individual gene product as a "pixel" of information, which relates to the expression of that, and only that, gene. We teach herein, methods whereby the individual "pixels" of gene expression information can be
5 combined into a single gene transcript "image," in which each of the individual genes can be visualized simultaneously and allowing relationships between the gene pixels to be easily visualized and understood.

We further teach a new method which we call electronic
10 subtraction. Electronic subtraction will enable the gene researcher to turn a single image into a moving picture, one which describes the temporality or dynamics of gene expression, at the level of a cell or a whole tissue. It is that sense of "motion" of cellular machinery on the
15 scale of a cell or organ which constitutes the new invention herein. This constitutes a new view into the process of living cell physiology and one which holds great promise to unveil and discover new therapeutic and diagnostic approaches in medicine.

20 We teach another method which we call "electronic northern," which tracks the expression of a single gene across many types of cells and tissues.

Nucleic acids (DNA and RNA) carry within their sequence the hereditary information and are therefore the
25 prime molecules of life. Nucleic acids are found in all living organisms including bacteria, fungi, viruses, plants and animals. It is of interest to determine the relative abundance of different discrete nucleic acids in different cells, tissues and organisms over time under various
30 conditions, treatments and regimes.

All dividing cells in the human body contain the same set of 23 pairs of chromosomes. It is estimated that these autosomal and sex chromosomes encode approximately 100,000 genes. The differences among different types of cells are
35 believed to reflect the differential expression of the 100,000 or so genes. Fundamental questions of biology could be answered by understanding which genes are transcribed and knowing the relative abundance of transcripts in different cells.

Previously, the art has only provided for the analysis of a few known genes at a time by standard molecular biology techniques such as PCR, northern blot analysis, or other types of DNA probe analysis such as in situ hybridization. Each of these methods allows one to analyze the transcription of only known genes and/or small numbers of genes at a time. Nucl. Acids Res. 19, 7097-7104 (1991); Nucl. Acids Res. 18, 4833-42 (1990); Nucl. Acids Res. 18, 2789-92 (1989); European J. Neuroscience 2, 1063-1073 (1990); Analytical Biochem. 187, 364-73 (1990); Genet. Annals Techn. Appl. 7, 64-70 (1990); GATA 8(4), 129-33 (1991); Proc. Natl. Acad. Sci. USA 85, 1696-1700 (1988); Nucl. Acids Res. 19, 1954 (1991); Proc. Natl. Acad. Sci. USA 88, 1943-47 (1991); Nucl. Acids Res. 19, 6123-27 (1991); Proc. Natl. Acad. Sci. USA 85, 5738-42 (1988); Nucl. Acids Res. 16, 10937 (1988).

Studies of the number and types of genes whose transcription is induced or otherwise regulated during cell processes such as activation, differentiation, aging, viral transformation, morphogenesis, and mitosis have been pursued for many years, using a variety of methodologies. One of the earliest methods was to isolate and analyze levels of the proteins in a cell, tissue, organ system, or even organisms both before and after the process of interest. One method of analyzing multiple proteins in a sample is using 2-dimensional gel electrophoresis, wherein proteins can be, in principle, identified and quantified as individual bands, and ultimately reduced to a discrete signal. At present, 2-dimensional analysis only resolves approximately 15% of the proteins. In order to positively analyze those bands which are resolved, each band must be excised from the membrane and subjected to protein sequence analysis using Edman degradation. Unfortunately, most of the bands were present in quantities too small to obtain a reliable sequence, and many of those bands contained more than one discrete protein. An additional difficulty is that many of the proteins were blocked at the amino-terminus, further complicating the sequencing process.

Analyzing differentiation at the gene transcription level has overcome many of these disadvantages and drawbacks, since the power of recombinant DNA technology allows amplification of signals containing very small amounts of material. The most common method, called "hybridization subtraction," involves isolation of mRNA from the biological specimen before (B) and after (A) the developmental process of interest, transcribing one set of mRNA into cDNA, subtracting specimen B from specimen A (mRNA from cDNA) by hybridization, and constructing a cDNA library from the non-hybridizing mRNA fraction. Many different groups have used this strategy successfully, and a variety of procedures have been published and improved upon using this same basic scheme. Nucl. Acids Res. 19, 7097-7104 (1991); Nucl. Acids Res. 18, 4833-42 (1990); Nucl. Acids Res. 18, 2789-92 (1989); European J. Neuroscience 2, 1063-1073 (1990); Analytical Biochem. 187, 364-73 (1990); Genet. Annals Techn. Appl. 7, 64-70 (1990); GATA 8(4), 129-33 (1991); Proc. Natl. Acad. Sci. USA 85, 1696-1700 (1988); Nucl. Acids Res. 19, 1954 (1991); Proc. Natl. Acad. Sci. USA 88, 1943-47 (1991); Nucl. Acids Res. 19, 6123-27 (1991); Proc. Natl. Acad. Sci. USA 85, 5738-42 (1988); Nucl. Acids Res. 16, 10937 (1988).

Although each of these techniques have particular strengths and weaknesses, there are still some limitations and undesirable aspects of these methods: First, the time and effort required to construct such libraries is quite large. Typically, a trained molecular biologist might expect construction and characterization of such a library to require 3 to 6 months, depending on the level of skill, experience, and luck. Second, the resulting subtraction libraries are typically inferior to the libraries constructed by standard methodology. A typical conventional cDNA library should have a clone complexity of at least 10^6 clones, and an average insert size of 1-3 kB. In contrast, subtracted libraries can have complexities of 10^2 or 10^3 and average insert sizes of 0.2 kB. Therefore, there can be a significant loss of clone and sequence information associated with such libraries. Third, this

approach allows the researcher to capture only the genes induced in specimen A relative to specimen B, not vice-versa, nor does it easily allow comparison to a third specimen of interest (C). Fourth, this approach requires
5 very large amounts (hundreds of micrograms) of "driver" mRNA (specimen B), which significantly limits the number and type of subtractions that are possible since many tissues and cells are very difficult to obtain in large quantities.

10 Fifth, the resolution of the subtraction is dependent upon the physical properties of DNA:DNA or RNA:DNA hybridization. The ability of a given sequence to find a hybridization match is dependent on its unique CoT value. The CoT value is a function of the number of copies
15 (concentration) of the particular sequence, multiplied by the time of hybridization. It follows that for sequences which are abundant, hybridization events will occur very rapidly (low CoT value), while rare sequences will form duplexes at very high CoT values. CoT values which allow
20 such rare sequences to form duplexes and therefore be effectively selected are difficult to achieve in a convenient time frame. Therefore, hybridization subtraction is simply not a useful technique with which to study relative levels of rare mRNA species. Sixth, this
25 problem is further complicated by the fact that duplex formation is also dependent on the nucleotide base composition for a given sequence. Those sequences rich in G + C form stronger duplexes than those with high contents of A + T. Therefore, the former sequences will tend to be
30 removed selectively by hybridization subtraction. Seventh, it is possible that hybridization between nonexact matches can occur. When this happens, the expression of a homologous gene may "mask" expression of a gene of interest, artificially skewing the results for that
35 particular gene.

Matsubara and Okubo proposed using partial cDNA sequences to establish expression profiles of genes which could be used in functional analyses of the human genome. Matsubara and Okubo warned against using random priming, as

it creates multiple unique DNA fragments from individual mRNAs and may thus skew the analysis of the number of particular mRNAs per library. They sequenced randomly selected members from a 3'-directed cDNA library and
5 established the frequency of appearance of the various ESTs. They proposed comparing lists of ESTs from various cell types to classify genes. Genes expressed in many different cell types were labeled housekeepers and those selectively expressed in certain cells were labeled cell-
10 specific genes, even in the absence of the full sequence of the gene or the biological activity of the gene product.

The present invention avoids the drawbacks of the prior art by providing a method to quantify the relative abundance of multiple gene transcripts in a given
15 biological specimen by the use of high-throughput sequence-specific analysis of individual RNAs and/or their corresponding cDNAs.

The present invention offers several advantages over current protein discovery methods which attempt to isolate
20 individual proteins based upon biological effects. The method of the instant invention provides for detailed diagnostic comparisons of cell profiles revealing numerous changes in the expression of individual transcripts.

The instant invention provides several advantages over
25 current subtraction methods including a more complex library analysis (10^6 to 10^7 clones as compared to 10^3 clones) which allows identification of low abundance messages as well as enabling the identification of messages which either increase or decrease in abundance. These
30 large libraries are very routine to make in contrast to the libraries of previous methods. In addition, homologues can easily be distinguished with the method of the instant invention.

This method is very convenient because it organizes a
35 large quantity of data into a comprehensible, digestible format. The most significant differences are highlighted by electronic subtraction. In depth analyses are made more convenient.

The present invention provides several advantages over previous methods of electronic analysis of cDNA. The method is particularly powerful when more than 100 and preferably more than 1,000 gene transcripts are analyzed.

5 In such a case, new low-frequency transcripts are discovered and tissue typed.

High resolution analysis of gene expression can be used directly as a diagnostic profile or to identify disease-specific genes for the development of more classic
10 diagnostic approaches.

This process is defined as gene transcript frequency analysis. The resulting quantitative analysis of the gene transcripts is defined as comparative gene transcript analysis.

15

3. SUMMARY OF THE INVENTION

The invention is a method of analyzing a specimen containing gene transcripts comprising the steps of (a) producing a library of biological sequences; (b) generating a set of transcript sequences, where each of the transcript
20 sequences in said set is indicative of a different one of the biological sequences of the library; (c) processing the transcript sequences in a programmed computer (in which a database of reference transcript sequences indicative of reference sequences is stored), to generate an identified
25 sequence value for each of the transcript sequences, where each said identified sequence value is indicative of sequence annotation and a degree of match between one of the biological sequences of the library and at least one of the reference sequences; and (d) processing each said
30 identified sequence value to generate final data values indicative of the number of times each identified sequence value is present in the library.

The invention also includes a method of comparing two specimens containing gene transcripts. The first specimen
35 is processed as described above. The second specimen is used to produce a second library of biological sequences, which is used to generate a second set of transcript sequences, where each of the transcript sequences in the

second set is indicative of one of the biological sequences of the second library. Then the second set of transcript sequences is processed in a programmed computer to generate a second set of identified sequence values, namely the
5 further identified sequence values, each of which is indicative of a sequence annotation and includes a degree of match between one of the biological sequences of the second library and at least one of the reference sequences. The further identified sequence values are processed to
10 generate further final data values indicative of the number of times each further identified sequence value is present in the second library. The final data values from the first specimen and the further identified sequence values from the second specimen are processed to generate ratios
15 of transcript sequences, which indicate the differences in the number of gene transcripts between the two specimens.

In a further embodiment, the method includes quantifying the relative abundance of mRNA in a biological specimen by (a) isolating a population of mRNA transcripts
20 from a biological specimen; (b) identifying genes from which the mRNA was transcribed by a sequence-specific method; (c) determining the numbers of mRNA transcripts corresponding to each of the genes; and (d) using the mRNA transcript numbers to determine the relative abundance of
25 mRNA transcripts within the population of mRNA transcripts.

Also disclosed is a method of producing a gene transcript image analysis by first obtaining a mixture of mRNA, from which cDNA copies are made. The cDNA is inserted into a suitable vector which is used to transfect
30 suitable host strain cells which are plated out and permitted to grow into clones, each clone representing a unique mRNA. A representative population of clones transfected with cDNA is isolated. Each clone in the population is identified by a sequence-specific method
35 which identifies the gene from which the unique mRNA was transcribed. The number of times each gene is identified to a clone is determined to evaluate gene transcript abundance. The genes and their abundances are listed in order of abundance to produce a gene transcript image.

In a further embodiment, the relative abundance of the gene transcripts in one cell type or tissue is compared with the relative abundance of gene transcript numbers in a second cell type or tissue in order to identify the differences and similarities.

In a further embodiment, the method includes a system for analyzing a library of biological sequences including a means for receiving a set of transcript sequences, where each of the transcript sequences is indicative of a different one of the biological sequences of the library; and a means for processing the transcript sequences in a computer system in which a database of reference transcript sequences indicative of reference sequences is stored, wherein the computer is programmed with software for generating an identified sequence value for each of the transcript sequences, where each said identified sequence value is indicative of a sequence annotation and the degree of match between a different one of the biological sequences of the library and at least one of the reference sequences, and for processing each said identified sequence value to generate final data values indicative of the number of times each identified sequence value is present in the library.

In essence, the invention is a method and system for quantifying the relative abundance of gene transcripts in a biological specimen. The invention provides a method for comparing the gene transcript image from two or more different biological specimens in order to distinguish between the two specimens and identify one or more genes which are differentially expressed between the two specimens. Thus, this gene transcript image and its comparison can be used as a diagnostic. One embodiment of the method generates high-throughput sequence-specific analysis of multiple RNAs or their corresponding cDNAs: a gene transcript image. Another embodiment of the method produces the gene transcript imaging analysis by the use of high-throughput cDNA sequence analysis. In addition, two or more gene transcript images can be compared and used to detect or diagnose a particular biological state, disease,

or condition which is correlated to the relative abundance of gene transcripts in a given cell or population of cells.

4. DESCRIPTION OF THE TABLES AND DRAWINGS

4.1. TABLES

5 Table 1 presents a detailed explanation of the letter codes utilized in Tables 2-5.

Table 2 lists the one hundred most common gene transcripts. It is a partial list of isolates from the HUVEC cDNA library prepared and sequenced as described
10 below. The left-hand column refers to the sequence's order of abundance in this table. The next column labeled "number" is the clone number of the first HUVEC sequence identification reference matching the sequence in the "entry" column number. Isolates that have not been
15 sequenced are not present in Table 2. The next column, labeled "N", indicates the total number of cDNAs which have the same degree of match with the sequence of the reference transcript in the "entry" column.

 The column labeled "entry" gives the NIH GENBANK locus
20 name, which corresponds to the library sequence numbers. The "s" column indicates in a few cases the species of the reference sequence. The code for column "s" is given in Table 1. The column labeled "descriptor" provides a plain English explanation of the identity of the sequence
25 corresponding to the NIH GENBANK locus name in the "entry" column.

Table 3 is a comparison of the top fifteen most abundant gene transcripts in normal monocytes and activated macrophage cells.

30 Table 4 is a detailed summary of library subtraction analysis summary comparing the THP-1 and human macrophage cDNA sequences. In Table 4, the same code as in Table 2 is used. Additional columns are for "bgfreq" (abundance number in the subtractant library), "rfend" (abundance
35 number in the target library) and "ratio" (the target abundance number divided by the subtractant abundance number). As is clear from perusal of the table, when the abundance number in the subtractant library is "0", the

target abundance number is divided by 0.05. This is a way of obtaining a result (not possible dividing by 0) and distinguishing the result from ratios of subtractant numbers of 1.

5 Table 5 is the computer program, written in source code, for generating gene transcript subtraction profiles.

Table 6 is a partial listing of database entries used in the electronic northern blot analysis as provided by the present invention.

10

4.2. BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 is a chart summarizing data collected and stored regarding the library construction portion of sequence preparation and analysis.

15 Figure 2 is a diagram representing the sequence of operations performed by "abundance sort" software in a class of preferred embodiments of the inventive method.

Figure 3 is a block diagram of a preferred embodiment of the system of the invention.

20 Figure 4 is a more detailed block diagram of the bioinformatics process from new sequence (that has already been sequenced but not identified) to printout of the transcript imaging analysis and the provision of database subscriptions.

25 5. DETAILED DESCRIPTION OF THE INVENTION

 The present invention provides a method to compare the relative abundance of gene transcripts in different biological specimens by the use of high-throughput sequence-specific analysis of individual RNAs or their
30 corresponding cDNAs (or alternatively, of data representing other biological sequences). This process is denoted herein as gene transcript imaging. The quantitative analysis of the relative abundance for a set of gene transcripts is denoted herein as "gene transcript image
35 analysis" or "gene transcript frequency analysis". The present invention allows one to obtain a profile for gene transcription in any given population of cells or tissue from any type of organism. The invention can be applied to

obtain a profile of a specimen consisting of a single cell (or clones of a single cell), or of many cells, or of tissue more complex than a single cell and containing multiple cell types, such as liver.

5 The invention has significant advantages in the fields of diagnostics, toxicology and pharmacology, to name a few. A highly sophisticated diagnostic test can be performed on the ill patient in whom a diagnosis has not been made. A biological specimen consisting of the patient's fluids or
10 tissues is obtained, and the gene transcripts are isolated and expanded to the extent necessary to determine their identity. Optionally, the gene transcripts can be converted to cDNA. A sampling of the gene transcripts are subjected to sequence-specific analysis and quantified.
15 These gene transcript sequence abundances are compared against reference database sequence abundances including normal data sets for diseased and healthy patients. The patient has the disease(s) with which the patient's data set most closely correlates.

20 For example, gene transcript frequency analysis can be used to differentiate normal cells or tissues from diseased cells or tissues, just as it highlights differences between normal monocytes and activated macrophages in Table 3.

 In toxicology, a fundamental question is which tests
25 are most effective in predicting or detecting a toxic effect. Gene transcript imaging provides highly detailed information on the cell and tissue environment, some of which would not be obvious in conventional, less detailed screening methods. The gene transcript image is a more
30 powerful method to predict drug toxicity and efficacy. Similar benefits accrue in the use of this tool in pharmacology. The gene transcript image can be used selectively to look at protein categories which are expected to be affected, for example, enzymes which
35 detoxify toxins.

 In an alternative embodiment, comparative gene transcript frequency analysis is used to differentiate between cancer cells which respond to anti-cancer agents and those which do not respond. Examples of anti-cancer

agents are tamoxifen, vincristine, vinblastine, podophyllotoxins, etoposide, teniposide, cisplatin, biologic response modifiers such as interferon, Il-2, GM-CSF, enzymes, hormones and the like. This method also
5 provides a means for sorting the gene transcripts by functional category. In the case of cancer cells, transcription factors or other essential regulatory molecules are very important categories to analyze across different libraries.

10 In yet another embodiment, comparative gene transcript frequency analysis is used to differentiate between control liver cells and liver cells isolated from patients treated with experimental drugs like FIAU to distinguish between pathology caused by the underlying disease and that caused
15 by the drug.

In yet another embodiment, comparative gene transcript frequency analysis is used to differentiate between brain tissue from patients treated and untreated with lithium.

In a further embodiment, comparative gene transcript
20 frequency analysis is used to differentiate between cyclosporin and FK506-treated cells and normal cells.

In a further embodiment, comparative gene transcript frequency analysis is used to differentiate between virally infected (including HIV-infected) human cells and
25 uninfected human cells. Gene transcript frequency analysis is also used to rapidly survey gene transcripts in HIV-resistant, HIV-infected, and HIV-sensitive cells. Comparison of gene transcript abundance will indicate the success of treatment and/or new avenues to study.

30 In a further embodiment, comparative gene transcript frequency analysis is used to differentiate between bronchial lavage fluids from healthy and unhealthy patients with a variety of ailments.

In a further embodiment, comparative gene transcript
35 frequency analysis is used to differentiate between cell, plant, microbial and animal mutants and wild-type species. In addition, the transcript abundance program is adapted to permit the scientist to evaluate the transcription of one gene in many different tissues. Such comparisons could

identify deletion mutants which do not produce a gene product and point mutants which produce a less abundant or otherwise different message. Such mutations can affect basic biochemical and pharmacological processes, such as mineral nutrition and metabolism, and can be isolated by means known to those skilled in the art. Thus, crops with improved yields, pest resistance and other factors can be developed.

In a further embodiment, comparative gene transcript frequency analysis is used for an interspecies comparative analysis which would allow for the selection of better pharmacologic animal models. In this embodiment, humans and other animals (such as a mouse), or their cultured cells are treated with a specific test agent. The relative sequence abundance of each cDNA population is determined. If the animal test system is a good model, homologous genes in the animal cDNA population should change expression similarly to those in human cells. If side effects are detected with the drug, a detailed transcript abundance analysis will be performed to survey gene transcript changes. Models will then be evaluated by comparing basic physiological changes.

In a further embodiment, comparative gene transcript frequency analysis is used in a clinical setting to give a highly detailed gene transcript profile of a patient's cells or tissue (for example, a blood sample). In particular, gene transcript frequency analysis is used to give a high resolution gene expression profile of a diseased state or condition.

In the preferred embodiment, the method utilizes high-throughput cDNA sequencing to identify specific transcripts of interest. The generated cDNA and deduced amino acid sequences are then extensively compared with GENBANK and other sequence data banks as described below. The method offers several advantages over current protein discovery by two-dimensional gel methods which try to identify individual proteins involved in a particular biological effect. Here, detailed comparisons of profiles of activated and inactive cells reveal numerous changes in

the expression of individual transcripts. After it is determined if the sequence is an "exact" match, similar or a non-match, the sequence is entered into a database. Next, the numbers of copies of cDNA corresponding to each gene are tabulated. Although this can be done slowly and arduously, if at all, by human hand from a printout of all entries, a computer program is a useful and rapid way to tabulate this information. The numbers of cDNA copies (optionally divided by the total number of sequences in the data set) provides a picture of the relative abundance of transcripts for each corresponding gene. The list of represented genes can then be sorted by abundance in the cDNA population. A multitude of additional types of comparisons or dimensions are possible and are exemplified below.

An alternate method of producing a gene transcript image includes the steps of obtaining a mixture of test mRNA and providing a representative array of unique probes whose sequences are complementary to at least some of the test mRNAs. Next, a fixed amount of the test mRNA is added to the arrayed probes. The test mRNA is incubated with the probes for a sufficient time to allow hybrids of the test mRNA and probes to form. The mRNA-probe hybrids are detected and the quantity determined. The hybrids are identified by their location in the probe array. The quantity of each hybrid is summed to give a population number. Each hybrid quantity is divided by the population number to provide a set of relative abundance data termed a gene transcript image analysis.

30

6. EXAMPLES

The examples below are provided to illustrate the subject invention. These examples are provided by way of illustration and are not included for the purpose of limiting the invention.

35

6.1. TISSUE SOURCES AND CELL LINES

For analysis with the computer program claimed herein, biological sequences can be obtained from virtually any

source. Most popular are tissues obtained from the human body. Tissues can be obtained from any organ of the body, any age donor, any abnormality or any immortalized cell line. Immortal cell lines may be preferred in some instances because of their purity of cell type; other tissue samples invariably include mixed cell types. A special technique is available to take a single cell (for example, a brain cell) and harness the cellular machinery to grow up sufficient cDNA for sequencing by the techniques and analysis described herein (cf. U.S. Patent Nos. 5,021,335 and 5,168,038, which are incorporated by reference). The examples given herein utilized the following immortalized cell lines: monocyte-like U-937 cells, activated macrophage-like THP-1 cells, induced vascular endothelial cells (HUVEC cells) and mast cell-like HMC-1 cells.

The U-937 cell line is a human histiocytic lymphoma cell line with monocyte characteristics, established from malignant cells obtained from the pleural effusion of a patient with diffuse histiocytic lymphoma (Sundstrom, C. and Nilsson, K. (1976) Int. J. Cancer 17:565). U-937 is one of only a few human cell lines with the morphology, cytochemistry, surface receptors and monocyte-like characteristics of histiocytic cells. These cells can be induced to terminal monocytic differentiation and will express new cell surface molecules when activated with supernatants from human mixed lymphocyte cultures. Upon this type of in vitro activation, the cells undergo morphological and functional changes, including augmentation of antibody-dependent cellular cytotoxicity (ADCC) against erythroid and tumor target cells (one of the principal functions of macrophages). Activation of U-937 cells with phorbol 12-myristate 13-acetate (PMA) in vitro stimulates the production of several compounds, including prostaglandins, leukotrienes and platelet-activating factor (PAF), which are potent inflammatory mediators. Thus, U-937 is a cell line that is well suited for the identification and isolation of gene transcripts associated with normal monocytes.

The HUVEC cell line is a normal, homogeneous, well characterized, early passage endothelial cell culture from human umbilical vein (Cell Systems Corp., 12815 NE 124th Street, Kirkland, WA 98034). Only gene transcripts from induced, or treated, HUVEC cells were sequenced. One batch of 1×10^8 cells was treated for 5 hours with 1 U/ml rIL-1b and 100 ng/ml E.coli lipopolysaccharide (LPS) endotoxin prior to harvesting. A separate batch of 2×10^8 cells was treated at confluence with 4 U/ml TNF and 2 U/ml interferon-gamma (IFN-gamma) prior to harvesting.

THP-1 is a human leukemic cell line with distinct monocytic characteristics. This cell line was derived from the blood of a 1-year-old boy with acute monocytic leukemia (Tsuchiya, S. et al. (1980) Int. J. Cancer: 171-76). The following cytological and cytochemical criteria were used to determine the monocytic nature of the cell line: 1) the presence of alpha-naphthyl butyrate esterase activity which could be inhibited by sodium fluoride; 2) the production of lysozyme; 3) the phagocytosis of latex particles and sensitized SRBC (sheep red blood cells); and 4) the ability of mitomycin C-treated THP-1 cells to activate T-lymphocytes following ConA (concanavalin A) treatment. Morphologically, the cytoplasm contained small azurophilic granules and the nucleus was indented and irregularly shaped with deep folds. The cell line had Fc and C3b receptors, probably functioning in phagocytosis. THP-1 cells treated with the tumor promoter 12-o-tetradecanoyl-phorbol-13 acetate (TPA) stop proliferating and differentiate into macrophage-like cells which mimic native monocyte-derived macrophages in several respects. Morphologically, as the cells change shape, the nucleus becomes more irregular and additional phagocytic vacuoles appear in the cytoplasm. The differentiated THP-1 cells also exhibit an increased adherence to tissue culture plastic.

HMC-1 cells (a human mast cell line) were established from the peripheral blood of a Mayo Clinic patient with mast cell leukemia (Leukemia Res. (1988) 12:345-55). The cultured cells looked similar to immature cloned murine

mast cells, contained histamine, and stained positively for chloroacetate esterase, amino caproate esterase, eosinophil major basic protein (MBP) and tryptase. The HMC-1 cells have, however, lost the ability to synthesize normal IgE
5 receptors. HMC-1 cells also possess a 10;16 translocation, present in cells initially collected by leukaphoresis from the patient and not an artifact of culturing. Thus, HMC-1 cells are a good model for mast cells.

6.2. CONSTRUCTION OF cDNA LIBRARIES

10 For inter-library comparisons, the libraries must be prepared in similar manners. Certain parameters appear to be particularly important to control. One such parameter is the method of isolating mRNA. It is important to use the same conditions to remove DNA and heterogeneous nuclear
15 RNA from comparison libraries. Size fractionation of cDNA must be carefully controlled. The same vector preferably should be used for preparing libraries to be compared. At the very least, the same type of vector (e.g., unidirectional vector) should be used to assure a valid
20 comparison. A unidirectional vector may be preferred in order to more easily analyze the output.

It is preferred to prime only with oligo dT unidirectional primer in order to obtain one only clone per mRNA transcript when obtaining cDNAs. However, it is
25 recognized that employing a mixture of oligo dT and random primers can also be advantageous because such a mixture results in more sequence diversity when gene discovery also is a goal. Similar effects can be obtained with DR2 (Clontech) and HXLOX (US Biochemical) and also vectors from
30 Invitrogen and Novagen. These vectors have two requirements. First, there must be primer sites for commercially available primers such as T3 or M13 reverse primers. Second, the vector must accept inserts up to 10 kB.

35 It also is important that the clones be randomly sampled, and that a significant population of clones is used. Data have been generated with 5,000 clones; however, if very rare genes are to be obtained and/or their relative

abundance determined, as many as 100,000 clones from a single library may need to be sampled. Size fractionation of cDNA also must be carefully controlled. Alternately, plaques can be selected, rather than clones.

5 Besides the Uni-ZAP™ vector system by Stratagene disclosed below, it is now believed that other similarly unidirectional vectors also can be used. For example, it is believed that such vectors include but are not limited to DR2 (Clontech), and HXLOX (U.S. Biochemical).

10 Preferably, the details of library construction (as shown in Figure 1) are collected and stored in a database for later retrieval relative to the sequences being compared. Fig. 1 shows important information regarding the library collaborator or cell or cDNA supplier,
15 pretreatment, biological source, culture, mRNA preparation and cDNA construction. Similarly detailed information about the other steps is beneficial in analyzing sequences and libraries in depth.

RNA must be harvested from cells and tissue samples
20 and cDNA libraries are subsequently constructed. cDNA libraries can be constructed according to techniques known in the art. (See, for example, Maniatis, T. et al. (1982) Molecular Cloning, Cold Spring Harbor Laboratory, New York). cDNA libraries may also be purchased. The U-937
25 cDNA library (catalog No. 937207) was obtained from Stratagene, Inc., 11099 M. Torrey Pines Rd., La Jolla, CA 92037.

The THP-1 cDNA library was custom constructed by Stratagene from THP-1 cells cultured 48 hours with 100 nm
30 TPA and 4 hours with 1 µg/ml LPS. The human mast cell HMC-1 cDNA library was also custom constructed by Stratagene from cultured HMC-1 cells. The HUVEC cDNA library was custom constructed by Stratagene from two batches of induced HUVEC cells which were separately processed.

35 Essentially, all the libraries were prepared in the same manner. First, poly(A+)RNA (mRNA) was purified. For the U-937 and HMC-1 RNA, cDNA synthesis was only primed with oligo dT. For the THP-1 and HUVEC RNA, cDNA synthesis was primed separately with both oligo dT and random

hexamers, and the two cDNA libraries were treated separately. Synthetic adaptor oligonucleotides were ligated onto cDNA ends enabling its insertion into the Uni-Zap™ vector system (Stratagene), allowing high efficiency
5 unidirectional (sense orientation) lambda library construction and the convenience of a plasmid system with blue-white color selection to detect clones with cDNA insertions. Finally, the two libraries were combined into a single library by mixing equal numbers of bacteriophage.
10 The libraries can be screened with either DNA probes or antibody probes and the pBluescript® phagemid (Stratagene) can be rapidly excised in vivo. The phagemid allows the use of a plasmid system for easy insert characterization, sequencing, site-directed mutagenesis,
15 the creation of unidirectional deletions and expression of fusion proteins. The custom-constructed library phage particles were infected into E. coli host strain XL1-Blue® (Stratagene), which has a high transformation efficiency, increasing the probability of obtaining rare, under-
20 represented clones in the cDNA library.

6.3. ISOLATION OF cDNA CLONES

The phagemid forms of individual cDNA clones were obtained by the in vivo excision process, in which the host bacterial strain was coinfectd with both the lambda
25 library phage and an f1 helper phage. Proteins derived from both the library-containing phage and the helper phage nicked the lambda DNA, initiated new DNA synthesis from defined sequences on the lambda target DNA and created a smaller, single stranded circular phagemid DNA molecule
30 that included all DNA sequences of the pBluescript® plasmid and the cDNA insert. The phagemid DNA was secreted from the cells and purified, then used to re-infect fresh host cells, where the double stranded phagemid DNA was produced. Because the phagemid carries the gene for beta-lactamase,
35 the newly-transformed bacteria are selected on medium containing ampicillin.

Phagemid DNA was purified using the Magic Minipreps™ DNA Purification System (Promega catalogue #A7100. Promega

Corp., 2800 Woods Hollow Rd., Madison, WI 53711). This small-scale process provides a simple and reliable method for lysing the bacterial cells and rapidly isolating purified phagemid DNA using a proprietary DNA-binding resin. The DNA was eluted from the purification resin already prepared for DNA sequencing and other analytical manipulations.

Phagemid DNA was also purified using the QIAwell-8 Plasmid Purification System from QIAGEN® DNA Purification System (QIAGEN Inc., 9259 Eton Ave., Chatsworth, CA 91311). This product line provides a convenient, rapid and reliable high-throughput method for lysing the bacterial cells and isolating highly purified phagemid DNA using QIAGEN anion-exchange resin particles with EMPORE™ membrane technology from 3M in a multiwell format. The DNA was eluted from the purification resin already prepared for DNA sequencing and other analytical manipulations.

An alternate method of purifying phagemid has recently become available. It utilizes the Miniprep Kit (Catalog No. 77468, available from Advanced Genetic Technologies Corp., 19212 Orbit Drive, Gaithersburg, Maryland). This kit is in the 96-well format and provides enough reagents for 960 purifications. Each kit is provided with a recommended protocol, which has been employed except for the following changes. First, the 96 wells are each filled with only 1 ml of sterile terrific broth with carbenicillin at 25 mg/L and glycerol at 0.4%. After the wells are inoculated, the bacteria are cultured for 24 hours and lysed with 60 µl of lysis buffer. A centrifugation step (2900 rpm for 5 minutes) is performed before the contents of the block are added to the primary filter plate. The optional step of adding isopropanol to TRIS buffer is not routinely performed. After the last step in the protocol, samples are transferred to a Beckman 96-well block for storage.

Another new DNA purification system is the WIZARD™ product line which is available from Promega (catalog No. A7071) and may be adaptable to the 96-well format.

6.4. SEQUENCING OF cDNA CLONES

The cDNA inserts from random isolates of the U-937 and THP-1 libraries were sequenced in part. Methods for DNA sequencing are well known in the art. Conventional enzymatic methods employ DNA polymerase Klenow fragment, Sequenase™ or Taq polymerase to extend DNA chains from an oligonucleotide primer annealed to the DNA template of interest. Methods have been developed for the use of both single- and double-stranded templates. The chain termination reaction products are usually electrophoresed on urea-acrylamide gels and are detected either by autoradiography (for radionuclide-labeled precursors) or by fluorescence (for fluorescent-labeled precursors). Recent improvements in mechanized reaction preparation, sequencing and analysis using the fluorescent detection method have permitted expansion in the number of sequences that can be determined per day (such as the Applied Biosystems 373 and 377 DNA sequencer, Catalyst 800). Currently with the system as described, read lengths range from 250 to 400 bases and are clone dependent. Read length also varies with the length of time the gel is run. In general, the shorter runs tend to truncate the sequence. A minimum of only about 25 to 50 bases is necessary to establish the identification and degree of homology of the sequence. Gene transcript imaging can be used with any sequence-specific method, including, but not limited to hybridization, mass spectroscopy, capillary electrophoresis and 505 gel electrophoresis.

30 6.5. HOMOLOGY SEARCHING OF cDNA CLONE AND DEDUCED PROTEIN (and Subsequent Steps)

Using the nucleotide sequences derived from the cDNA clones as query sequences (sequences of a Sequence Listing), databases containing previously identified sequences are searched for areas of homology (similarity). Examples of such databases include Genbank and EMBL. We next describe examples of two homology search algorithms that can be used, and then describe the subsequent computer-implemented steps to be performed in accordance with preferred embodiments of the invention.

In the following description of the computer-implemented steps of the invention, the word "library" denotes a set (or population) of biological specimen nucleic acid sequences. A "library" can consist of cDNA sequences, RNA sequences, or the like, which characterize a biological specimen. The biological specimen can consist of cells of a single human cell type (or can be any of the other above-mentioned types of specimens). We contemplate that the sequences in a library have been determined so as to accurately represent or characterize a biological specimen (for example, they can consist of representative cDNA sequences from clones of RNA taken from a single human cell).

In the following description of the computer-implemented steps of the invention, the expression "database" denotes a set of stored data which represent a collection of sequences, which in turn represent a collection of biological reference materials. For example, a database can consist of data representing many stored cDNA sequences which are in turn representative of human cells infected with various viruses, cells of humans of various ages, cells from different mammalian species, and so on.

In preferred embodiments, the invention employs a computer programmed with software (to be described) for performing the following steps:

(a) processing data indicative of a library of cDNA sequences (generated as a result of high-throughput cDNA sequencing or other method) to determine whether each sequence in the library matches a DNA sequence of a reference database of DNA sequences (and if so, identifying the reference database entry which matches the sequence and indicating the degree of match between the reference sequence and the library sequence) and assigning an identified sequence value based on the sequence annotation and degree of match to each of the sequences in the library;

(b) for some or all entries of the database, tabulating the number of matching identified sequence

values in the library (Although this can be done by human hand from a printout of all entries, we prefer to perform this step using computer software to be described below.), thereby generating a set of final data values or "abundance numbers"; and

(c) if the libraries are different sizes, dividing each abundance number by the total number of sequences in the library, to obtain a relative abundance number for each identified sequence value (i.e., a relative abundance of each gene transcript).

The list of identified sequence values (or genes corresponding thereto) can then be sorted by abundance in the cDNA population. A multitude of additional types of comparisons or dimensions are possible.

For example (to be described below in greater detail), steps (a) and (b) can be repeated for two different libraries (sometimes referred to as a "target" library and a "subtractant" library). Then, for each identified sequence value (or gene transcript), a "ratio" value is obtained by dividing the abundance number (for that identified sequence value) for the target library, by the abundance number (for that identified sequence value) for the subtractant library.

In fact, subtraction may be carried out on multiple libraries. It is possible to add the transcripts from several libraries (for example, three) and then to divide them by another set of transcripts from multiple libraries (again, for example, three). Notation for this operation may be abbreviated as $(A+B+C) / (D+E+F)$, where the capital letters each indicate an entire library. Optionally the abundance numbers of transcripts in the summed libraries may be divided by the total sample size before subtraction.

Unlike standard hybridization technology which permits a single subtraction of two libraries, once one has processed a set or library transcript sequences and stored them in the computer, any number of subtractions can be performed on the library. For example, by this method, ratio values can be obtained by dividing relative abundance

values in a first library by corresponding values in a second library and vice versa.

In variations on step (a), the library consists of nucleotide sequences derived from cDNA clones. Examples of
5 databases which can be searched for areas of homology (similarity) in step (a) include the commercially available databases known as Genbank (NIH) EMBL (European Molecular Biology Labs, Germany), and GENESEQ (Intelligenetics, Mountain View, California).

10 One homology search algorithm which can be used to implement step (a) is the algorithm described in the paper by D.J. Lipman and W.R. Pearson, entitled "Rapid and Sensitive Protein Similarity Searches," Science, 227:1435 (1985). In this algorithm, the homologous regions are
15 searched in a two-step manner. In the first step, the highest homologous regions are determined by calculating a matching score using a homology score table. The parameter "Ktup" is used in this step to establish the minimum window size to be shifted for comparing two sequences. Ktup also
20 sets the number of bases that must match to extract the highest homologous region among the sequences. In this step, no insertions or deletions are applied and the homology is displayed as an initial (INIT) value.

In the second step, the homologous regions are aligned
25 to obtain the highest matching score by inserting a gap in order to add a probable deleted portion. The matching score obtained in the first step is recalculated using the homology score Table and the insertion score Table to an optimized (OPT) value in the final output.

30 DNA homologies between two sequences can be examined graphically using the Harr method of constructing dot matrix homology plots (Needleman, S.B. and Wunsch, C.O., J. Mom. Biol 48:443 (1970)). This method produces a two-dimensional plot which can be useful in determining
35 regions of homology versus regions of repetition.

However, in a class of preferred embodiments, step (a) is implemented by processing the library data in the commercially available computer program known as the INHERIT 670 Sequence Analysis System, available from

Applied Biosystems Inc. (Foster City, California), including the software known as the Factura software (also available from Applied Biosystems Inc.). The Factura program preprocesses each library sequence to "edit out" portions thereof which are not likely to be of interest, such as the vector used to prepare the library. Additional sequences which can be edited out or masked (ignored by the search tools) include but are not limited to the polyA tail and repetitive GAG and CCC sequences. A low-end search program can be written to mask out such "low-information" sequences, or programs such as BLAST can ignore the low-information sequences.

In the algorithm implemented by the INHERIT 670 Sequence Analysis System, the Pattern Specification Language (developed by TRW Inc.) is used to determine regions of homology. "There are three parameters that determine how INHERIT analysis runs sequence comparisons: window size, window offset and error tolerance. Window size specifies the length of the segments into which the query sequence is subdivided. Window offset specifies where to start the next segment [to be compared], counting from the beginning of the previous segment. Error tolerance specifies the total number of insertions, deletions and/or substitutions that are tolerated over the specified word length. Error tolerance may be set to any integer between 0 and 6. The default settings are window tolerance=20, window offset=10 and error tolerance=3." INHERIT Analysis Users Manual, pp.2-15. Version 1.0, Applied Biosystems, Inc., October 1991.

Using a combination of these three parameters, a database (such as a DNA database) can be searched for sequences containing regions of homology and the appropriate sequences are scored with an initial value. Subsequently, these homologous regions are examined using dot matrix homology plots to determine regions of homology versus regions of repetition. Smith-Waterman alignments can be used to display the results of the homology search. The INHERIT software can be executed by a Sun computer system programmed with the UNIX operating system.

Search alternatives to INHERIT include the BLAST program, GCG (available from the Genetics Computer Group, WI) and the Dasher program (Temple Smith, Boston University, Boston, MA). Nucleotide sequences can be
5 searched against Genbank, EMBL or custom databases such as GENESEQ (available from Intelligenetics, Mountain View, CA) or other databases for genes. In addition, we have searched some sequences against our own in-house database.

In preferred embodiments, the transcript sequences are
10 analyzed by the INHERIT software for best conformance with a reference gene transcript to assign a sequence identifier and assigned the degree of homology, which together are the identified sequence value and are input into, and further processed by, a Macintosh personal computer (available from
15 Apple) programmed with an "abundance sort and subtraction analysis" computer program (to be described below).

Prior to the abundance sort and subtraction analysis program (also denoted as the "abundance sort" program), identified sequences from the cDNA clones are assigned
20 value (according to the parameters given above) by degree of match according to the following categories: "exact" matches (regions with a high degree of identity), homologous human matches (regions of high similarity, but not "exact" matches), homologous non-human matches (regions
25 of high similarity present in species other than human), or non matches (no significant regions of homology to previously identified nucleotide sequences stored in the form of the database). Alternately, the degree of match can be a numeric value as described below.

30 With reference again to the step of identifying matches between reference sequences and database entries, protein and peptide sequences can be deduced from the nucleic acid sequences. Using the deduced polypeptide sequence, the match identification can be performed in a
35 manner analogous to that done with cDNA sequences. A protein sequence is used as a query sequence and compared to the previously identified sequences contained in a database such as the Swiss/Prot, PIR and the NBRF Protein database to find homologous proteins. These proteins are

initially scored for homology using a homology score Table (Orcutt, B.C. and Dayoff, M.O. Scoring Matrices, PIR Report MAT - 0285 (February 1985)) resulting in an INIT score. The homologous regions are aligned to obtain the
5 highest matching scores by inserting a gap which adds a probable deleted portion. The matching score is recalculated using the homology score Table and the insertion score Table resulting in an optimized (OPT) score. Even in the absence of knowledge of the proper
10 reading frame of an isolated sequence, the above-described protein homology search may be performed by searching all 3 reading frames.

Peptide and protein sequence homologies can also be ascertained using the INHERIT 670 Sequence Analysis System
15 in an analogous way to that used in DNA sequence homologies. Pattern Specification Language and parameter windows are used to search protein databases for sequences containing regions of homology which are scored with an initial value. Subsequent display in a dot-matrix homology
20 plot shows regions of homology versus regions of repetition. Additional search tools that are available to use on pattern search databases include PLsearch Blocks (available from Henikoff & Henikoff, University of Washington, Seattle), Dasher and GCG. Pattern search
25 databases include, but are not limited to, Protein Blocks (available from Henikoff & Henikoff, University of Washington, Seattle), Brookhaven Protein (available from the Brookhaven National Laboratory, Brookhaven, MA), PROSITE (available from Amos Bairoch, University of Geneva, Switzerland), ProDom (available from Temple Smith, Boston
30 University), and PROTEIN MOTIF FINGERPRINT (available from University of Leeds, United Kingdom).

The ABI Assembler application software, part of the INHERIT DNA analysis system (available from Applied
35 Biosystems, Inc., Foster City, CA), can be employed to create and manage sequence assembly projects by assembling data from selected sequence fragments into a larger sequence. The Assembler software combines two advanced computer technologies which maximize the ability to

assemble sequenced DNA fragments into Assemblages, a special grouping of data where the relationships between sequences are shown by graphic overlap, alignment and statistical views. The process is based on the

5 Meyers-Kececiloglu model of fragment assembly (INHERIT™ Assembler User's Manual, Applied Biosystems, Inc., Foster City, CA), and uses graph theory as the foundation of a very rigorous multiple sequence alignment engine for assembling DNA sequence fragments. Other assembly programs

10 that can be used include MEGALIGN (available from DNASTAR Inc., Madison, WI), Dasher and STADEN (available from Roger Staden, Cambridge, England).

Next, with reference to Fig. 2, we describe in more detail the "abundance sort" program which implements above-

15 mentioned "step (b)" to tabulate the number of sequences of the library which match each database entry (the "abundance number" for each database entry).

Fig. 2 is a flow chart of a preferred embodiment of the abundance sort program. A source code listing of this

20 embodiment of the abundance sort program is set forth in Table 5. In the Table 5 implementation, the abundance sort program is written using the FoxBASE programming language commercially available from Microsoft Corporation. Although FoxBASE was the program chosen for the first

25 iteration of this technology, it should not be considered limiting. Many other programming languages, Sybase being a particularly desirable alternative, can also be used, as will be obvious to one with ordinary skill in the art. The subroutine names specified in Fig. 2 correspond to

30 subroutines listed in Table 5.

With reference again to Fig. 2, the "Identified Sequences" are transcript sequences representing each sequence of the library and a corresponding identification of the database entry (if any) which it matches. In other

35 words, the "Identified Sequences" are transcript sequences representing the output of above-discussed "step (a)."

Fig. 3 is a block diagram of a system for implementing the invention. The Fig. 3 system includes library generation unit 2 which generates a library and asserts an

output stream of transcript sequences indicative of the biological sequences comprising the library. Programmed processor 4 receives the data stream output from unit 2 and processes this data in accordance with above-discussed

5 "step (a)" to generate the Identified Sequences. Processor 4 can be a processor programmed with the commercially available computer program known as the INHERIT 670 Sequence Analysis System and the commercially available computer program known as the Factura program (both

10 available from Applied Biosystems Inc.) and with the UNIX operating system.

Still with reference to Fig. 3, the Identified Sequences are loaded into processor 6 which is programmed with the abundance sort program. Processor 6 generates the

15 Final Transcript sequences indicated in both Figs. 2 and 3. Fig. 4 shows a more detailed block diagram of a planned relational computer system, including various searching techniques which can be implemented, along with an assortment of databases to query against.

20 With reference to Fig. 2, the abundance sort program first performs an operation known as "Tempnum" on the Identified Sequences, to discard all of the Identified Sequences except those which match database entries of selected types. For example, the Tempnum process can

25 select Identified Sequences which represent matches of the following types with database entries (see above for definition): "exact" matches, human "homologous" matches, "other species" matches representing genes present in species other than human), "no" matches (no significant

30 regions of homology with database entries representing previously identified nucleotide sequences), "I" matches (Incyte for not previously known DNA sequences), or "X" matches (matches ESTs in reference database). This eliminates the U, S, M, V, A, R and D sequence (see Table 1

35 for definitions).

The identified sequence values selected during the "Tempnum" process then undergo a further selection (weeding out) operation known as "Tempred." This operation can, for

example, discard all identified sequence values representing matches with selected database entries.

The identified sequence values selected during the "Tempred" process are then classified according to library, during the "Tempdesig" operation. It is contemplated that the "Identified Sequences" can represent sequences from a single library, or from two or more libraries.

Consider first the case that the identified sequence values represent sequences from a single library. In this case, all the identified sequence values determined during "Tempred" undergo sorting in the "Templib" operation, further sorting in the "Libsort" operation, and finally additional sorting in the "Temptarsort" operation. For example, these three sorting operations can sort the identified sequences in order of decreasing "abundance number" (to generate a list of decreasing abundance numbers, each abundance number corresponding to a unique identified sequence entry, or several lists of decreasing abundance numbers, with the abundance numbers in each list corresponding to database entries of a selected type) with redundancies eliminated from each sorted list. In this case, the operation identified as "Cruncher" can be bypassed, so that the "Final Data" values are the organized transcript sequences produced during the "Temptarsort" operation.

We next consider the case that the transcript sequences produced during the "Tempred" operation represent sequences from two libraries (which we will denote the "target" library and the "subtractant" library). For example, the target library may consist of cDNA sequences from clones of a diseased cell, while the subtractant library may consist of cDNA sequences from clones of the diseased cell after treatment by exposure to a drug. For another example, the target library may consist of cDNA sequences from clones of a cell type from a young human, while the subtractant library may consist of cDNA sequences from clones of the same cell type from the same human at different ages.

In this case, the "Tempdesig" operation routes all transcript sequences representing the target library for processing in accordance with "Templib" (and then "Libsort" and "Temptarsort"), and routes all transcript sequences
5 representing the subtractant library for processing in accordance with "Tempsub" (and then "Subsort" and "Tempsubsort"). For example, the consecutive "Templib," "Libsort," and "Temptarsort" sorting operations sort identified sequences from the target library in order of
10 decreasing abundance number (to generate a list of decreasing abundance numbers, each abundance number corresponding to a database entry, or several lists of decreasing abundance numbers, with the abundance numbers in each list corresponding to database entries of a selected
15 type) with redundancies eliminated from each sorted list. The consecutive "Tempsub," "Subsort," and "Tempsubsort" sorting operations sort identified sequences from the subtractant library in order of decreasing abundance number (to generate a list of decreasing abundance numbers, each
20 abundance number corresponding to a database entry, or several lists of decreasing abundance numbers, with the abundance numbers in each list corresponding to database entries of a selected type) with redundancies eliminated from each sorted list.

25 The transcript sequences output from the "Temptarsort" operation typically represent sorted lists from which a histogram could be generated in which position along one (e.g., horizontal) axis indicates abundance number (of target library sequences), and position along another
30 (e.g., vertical) axis indicates identified sequence value (e.g., human or non-human gene type). Similarly, the transcript sequences output from the "Tempsubsort" operation typically represent sorted lists from which a histogram could be generated in which position along one
35 (e.g., horizontal) axis indicates abundance number (of subtractant library sequences), and position along another (e.g., vertical) axis indicates identified sequence value (e.g., human or non-human gene type).

The transcript sequences (sorted lists) output from the Tempsubsort and Temptarsort sorting operations are combined during the operation identified as "Cruncher." The "Cruncher" process identifies pairs of corresponding target and subtractant abundance numbers (both representing the same identified sequence value), and divides one by the other to generate a "ratio" value for each pair of corresponding abundance numbers, and then sorts the ratio values in order of decreasing ratio value. The data output from the "Cruncher" operation (the Final Transcript sequence in Fig. 2) is typically a sorted list from which a histogram could be generated in which position along one axis indicates the size of a ratio of abundance numbers (for corresponding identified sequence values from target and subtractant libraries) and position along another axis indicates identified sequence value (e.g., gene type).

Preferably, prior to obtaining a ratio between the two library abundance values, the Cruncher operation also divides each ratio value by the total number of sequences in one or both of the target and subtractant libraries. The resulting lists of "relative" ratio values generated by the Cruncher operation are useful for many medical, scientific, and industrial applications. Also preferably, the output of the Cruncher operation is a set of lists, each list representing a sequence of decreasing ratio values for a different selected subset (e.g. protein family) of database entries.

In one example, the abundance sort program of the invention tabulates for a library the numbers of mRNA transcripts corresponding to each gene identified in a database. These numbers are divided by the total number of clones sampled. The results of the division reflect the relative abundance of the mRNA transcripts in the cell type or tissue from which they were obtained. Obtaining this final data set is referred to herein as "gene transcript image analysis." The resulting subtracted data show exactly what proteins and genes are upregulated and downregulated in highly detailed complexity.

6.6. HUVEC cDNA LIBRARY

Table 2 is an abundance table listing the various gene transcripts in an induced HUVEC library. The transcripts are listed in order of decreasing abundance. This computerized sorting simplifies analysis of the tissue and speeds identification of significant new proteins which are specific to this cell type. This type of endothelial cell lines tissues of the cardiovascular system, and the more that is known about its composition, particularly in response to activation, the more choices of protein targets become available to affect in treating disorders of this tissue, such as the highly prevalent atherosclerosis.

6.7. MONOCYTE-CELL AND MAST-CELL cDNA LIBRARIES

Tables 3 and 4 show truncated comparisons of two libraries. In Tables 3 and 4 the "normal monocytes" are the HMC-1 cells, and the "activated macrophages" are the THP-1 cells pretreated with PMA and activated with LPS. Table 3 lists in descending order of abundance the most abundant gene transcripts for both cell types. With only 15 gene transcripts from each cell type, this table permits quick, qualitative comparison of the most common transcripts. This abundance sort, with its convenient side-by-side display, provides an immediately useful research tool. In this example, this research tool discloses that 1) only one of the top 15 activated macrophage transcripts is found in the top 15 normal monocyte gene transcripts (poly A binding protein); and 2) a new gene transcript (previously unreported in other databases) is relatively highly represented in activated macrophages but is not similarly prominent in normal macrophages. Such a research tool provides researchers with a short-cut to new proteins, such as receptors, cell-surface and intracellular signalling molecules, which can serve as drug targets in commercial drug screening programs. Such a tool could save considerable time over that consumed by a hit and miss discovery program aimed at identifying important proteins in and around cells, because those proteins carrying out everyday cellular functions and

represented as steady state mRNA are quickly eliminated from further characterization.

This illustrates how the gene transcript profiles change with altered cellular function. Those skilled in the art know that the biochemical composition of cells also changes with other functional changes such as cancer, including cancer's various stages, and exposure to toxicity. A gene transcript subtraction profile such as in Table 3 is useful as a first screening tool for such gene expression and protein studies.

6.8. SUBTRACTION ANALYSIS OF NORMAL MONOCYTE-CELL AND ACTIVATED MONOCYTE CELL cDNA LIBRARIES

Once the cDNA data are in the computer, the computer program as disclosed in Table 5 was used to obtain ratios of all the gene transcripts in the two libraries discussed in Example 6.7, and the gene transcripts were sorted by the descending values of their ratios. If a gene transcript is not represented in one library, that gene transcript's abundance is unknown but appears to be less than 1. As an approximation -- and to obtain a ratio, which would not be possible if the unrepresented gene were given an abundance of zero -- genes which are represented in only one of the two libraries are assigned an abundance of 1/2. Using 1/2 for unrepresented clones increases the relative importance of "turned-on" and "turned-off" genes, whose products would be drug candidates. The resulting print-out is called a subtraction table and is an extremely valuable screening method, as is shown by the following data.

Table 4 is a subtraction table, in which the normal monocyte library was electronically "subtracted" from the activated macrophage library. This table highlights most effectively the changes in abundance of the gene transcripts by activation of macrophages. Even among the first 20 gene transcripts listed, there are several unknown gene transcripts. Thus, electronic subtraction is a useful tool with which to assist researchers in identifying much more quickly the basic biochemical changes between two cell types. Such a tool can save universities and pharmaceutical companies which spend billions of dollars on

research valuable time and laboratory resources at the early discovery stage and can speed up the drug development cycle, which in turn permits researchers to set up drug screening programs much earlier. Thus, this research tool
5 provides a way to get new drugs to the public faster and more economically.

Also, such a subtraction table can be obtained for patient diagnosis. An individual patient sample (such as monocytes obtained from a biopsy or blood sample) can be
10 compared with data provided herein to diagnose conditions associated with macrophage activation.

Table 4 uncovered many new gene transcripts (labeled Incyte clones). Note that many genes are turned on in the activated macrophage (i.e., the monocyte had a 0 in the
15 bgfreq column). This screening method is superior to other screening techniques, such as the western blot, which are incapable of uncovering such a multitude of discrete new gene transcripts.

The subtraction-screening technique has also uncovered
20 a high number of cancer gene transcripts (oncogenes rho, ETS2, rab-2 ras, YPT1-related, and acute myeloid leukemia mRNA) in the activated macrophage. These transcripts may be attributed to the use of immortalized cell lines and are inherently interesting for that reason. This screening
25 technique offers a detailed picture of upregulated transcripts including oncogenes, which helps explain why anti-cancer drugs interfere with the patient's immunity mediated by activated macrophages. Armed with knowledge gained from this screening method, those skilled in the art
30 can set up more targeted, more effective drug screening programs to identify drugs which are differentially effective against 1) both relevant cancers and activated macrophage conditions with the same gene transcript
35 conditions.

Smooth muscle senescent protein (22 kd) was upregulated in the activated macrophage, which indicates that it is a candidate to block in controlling inflammation.

6.9. SUBTRACTION ANALYSIS OF NORMAL LIVER CELLS AND HEPATITIS INFECTED LIVER CELL cDNA LIBRARIES

In this example, rats are exposed to hepatitis virus and maintained in the colony until they show definite signs of hepatitis. Of the rats diagnosed with hepatitis, one half of the rats are treated with a new anti-hepatitis agent (AHA). Liver samples are obtained from all rats before exposure to the hepatitis virus and at the end of AHA treatment or no treatment. In addition, liver samples can be obtained from rats with hepatitis just prior to AHA treatment.

The liver tissue is treated as described in Examples 6.2 and 6.3 to obtain mRNA and subsequently to sequence cDNA. The cDNA from each sample are processed and analyzed for abundance according to the computer program in Table 5. The resulting gene transcript images of the cDNA provide detailed pictures of the baseline (control) for each animal and of the infected and/or treated state of the animals. cDNA data for a group of samples can be combined into a group summary gene transcript profile for all control samples, all samples from infected rats and all samples from AHA-treated rats.

Subtractions are performed between appropriate individual libraries and the grouped libraries. For individual animals, control and post-study samples can be subtracted. Also, if samples are obtained before and after AHA treatment, that data from individual animals and treatment groups can be subtracted. In addition, the data for all control samples can be pooled and averaged. The control average can be subtracted from averages of both post-study AHA and post-study non-AHA cDNA samples. If pre- and post-treatment samples are available, pre- and post-treatment samples can be compared individually (or electronically averaged) and subtracted.

These subtraction tables are used in two general ways. First, the differences are analyzed for gene transcripts which are associated with continuing hepatic deterioration or healing. The subtraction tables are tools to isolate the effects of the drug treatment from the underlying basic pathology of hepatitis. Because hepatitis affects many

parameters, additional liver toxicity has been difficult to detect with only blood tests for the usual enzymes. The gene transcript profile and subtraction provides a much more complex biochemical picture which researchers have
5 needed to analyze such difficult problems.

Second, the subtraction tables provide a tool for identifying clinical markers, individual proteins or other biochemical determinants which are used to predict and/or evaluate a clinical endpoint, such as disease, improvement
10 due to the drug, and even additional pathology due to the drug. The subtraction tables specifically highlight genes which are turned on or off. Thus, the subtraction tables provide a first screen for a set of gene transcript candidates for use as clinical markers. Subsequently,
15 electronic subtractions of additional cell and tissue libraries reveal which of the potential markers are in fact found in different cell and tissue libraries. Candidate gene transcripts found in additional libraries are removed from the set of potential clinical markers. Then, tests of
20 blood or other relevant samples which are known to lack and have the relevant condition are compared to validate the selection of the clinical marker. In this method, the particular physiologic function of the protein transcript need not be determined to qualify the gene transcript as a
25 clinical marker.

6.10. ELECTRONIC NORTHERN BLOT

One limitation of electronic subtraction is that it is difficult to compare more than a pair of images at once. Once particular individual gene products are identified as
30 relevant to further study (via electronic subtraction or other methods), it is useful to study the expression of single genes in a multitude of different tissues. In the lab, the technique of "Northern" blot hybridization is used for this purpose. In this technique, a single cDNA, or a
35 probe corresponding thereto, is labeled and then hybridized against a blot containing RNA samples prepared from a multitude of tissues or cell types. Upon autoradiography,

the pattern of expression of that particular gene, one at a time, can be quantitated in all the included samples.

In contrast, a further embodiment of this invention is the computerized form of this process, termed here

5 "electronic northern blot." In this variation, a single gene is queried for expression against a multitude of prepared and sequenced libraries present within the database. In this way, the pattern of expression of any single candidate gene can be examined instantaneously and

10 effortlessly. More candidate genes can thus be scanned, leading to more frequent and fruitfully relevant discoveries. The computer program included as Table 5 includes a program for performing this function, and Table 6 is a partial listing of entries of the database used in

15 the electronic northern blot analysis.

6.11. PHASE I CLINICAL TRIALS

Based on the establishment of safety and effectiveness in the above animal tests, Phase I clinical tests are undertaken. Normal patients are subjected to the usual

20 preliminary clinical laboratory tests. In addition, appropriate specimens are taken and subjected to gene transcript analysis. Additional patient specimens are taken at predetermined intervals during the test. The specimens are subjected to gene transcript analysis as

25 described above. In addition, the gene transcript changes noted in the earlier rat toxicity study are carefully evaluated as clinical markers in the followed patients. Changes in the gene transcript analyses are evaluated as

30 indicators of toxicity by correlation with clinical signs and symptoms and other laboratory results. In addition, subtraction is performed on individual patient specimens and on averaged patient specimens. The subtraction analysis highlights any toxicological changes in the treated patients. This is a highly refined determinant of

35 toxicity. The subtraction method also annotates clinical markers. Further subgroups can be analyzed by subtraction analysis, including, for example, 1) segregation by

occurrence and type of adverse effect; and 2) segregation by dosage.

6.12. GENE TRANSCRIPT IMAGING ANALYSIS IN CLINICAL STUDIES

A gene transcript imaging analysis (or multiple gene transcript imaging analyses) is a useful tool in other clinical studies. For example, the differences in gene transcript imaging analyses before and after treatment can be assessed for patients on placebo and drug treatment. This method also effectively screens for clinical markers to follow in clinical use of the drug.

6.13. COMPARATIVE GENE TRANSCRIPT ANALYSIS BETWEEN SPECIES

The subtraction method can be used to screen cDNA libraries from diverse sources. For example, the same cell types from different species can be compared by gene transcript analysis to screen for specific differences, such as in detoxification enzyme systems. Such testing aids in the selection and validation of an animal model for the commercial purpose of drug screening or toxicological testing of drugs intended for human or animal use. When the comparison between animals of different species is shown in columns for each species, we refer to this as an interspecies comparison, or zoo blot.

Embodiments of this invention may employ databases such as those written using the FoxBASE programming language commercially available from Microsoft Corporation. Other embodiments of the invention employ other databases, such as a random peptide database, a polymer database, a synthetic oligomer database, or a oligonucleotide database of the type described in U.S. Patent 5,270,170, issued December 14, 1993 to Cull, et al., PCT International Application Publication No. WO 9322684, published November 11, 1993, PCT International Application Publication No. WO 9306121, published April 1, 1993, or PCT International Application Publication No. WO 9119818, published December 26, 1991. These four references (whose text is incorporated herein by reference) include teaching which

may be applied in implementing such other embodiments of the present invention.

All references referred to in the preceding text are hereby expressly incorporated by reference herein.

- 5 Various modifications and variations of the described method and system of the invention will be apparent to those skilled in the art without departing from the scope and spirit of the invention. Although the invention has been described in connection with specific preferred
- 10 embodiments, it should be understood that the invention as claimed should not be unduly limited to such specific embodiments.

TABLE 1

Designations (D)	Distribution (F)	Localization (Z)	Function (R)
E = Exact	C = Non-specific	N = Nuclear	T = Translation
H = Homologous	P = Cell/tissue specific	C = Cytoplasmic	L = Protein processing
O = Other species	U = Unknown	K = Cytoskeleton	R = Ribosomal protein
N = No match		E = Cell surface	O = Oncogene
D = Noncoding gene		Z = Intracellular memb	G = GTP binding ptn
U = Nonreadable		M = Mitochondrial	V = Viral element
R = Repetitive DNA		S = Secreted	Y = Kinase/phosphatase
A = Poly-A only	Species (S)	U = Unknown	A = Tumor antigen related
V = Vector only	H = Human	X = Other	I = Binding proteins
M = Mitochondrial DNA	A = Ape		D = NA-binding /transcription
S = Skip	P = Pig		B = Surface molecule/receptor
I = Match Incyte clone	D = Dog		C = Ca ⁺⁺ binding protein
X = EST match	V = Bovine		S = Ligands/effectors
	B = Rabbit		H = Stress response protein
	R = Rat		E = Enzyme
Library (L)	M = Mouse	Status (I)	F = Ferroprotein
U = U937	S = Hamster	0 = No current interest	P = Protease/inhibitor
M = HMC	C = Chicken	1 = Do primary analysis	Z = Oxidative phosphorylation
T = THP-1	F = Amphibian	2 = Primary analysis done	Q = Sugar metabolism
H = HUVEC	I = Invertebrate	3 = Full length sequence	M = Amino acid metabolism
S = Spleen	Z = Protozoan	4 = Secondary analysis	N = Nucleic acid metabolism
L = Lung	G = Fungi	5 = Tissue northern	W = Lipid metabolism
Y = T & B cell		6 = Obtain full length	K = Structural
A = Adenoid			X = Other
			U = unknown

TABLE 2

Clone numbers 15000 through 20000

Libraries: HUVEC

Arranged by ABUNDANCE

Total clones analyzed: 5000

319 genes, for a total of 1713 Clones

	number	N	c	entry	s	descriptor
1	15365	67		HSRPL41		Riboptn L41
2	15004	65		NCY015004		INCYTE 015004
3	15638	63		NCY015638		INCYTE 015638
4	15390	50		NCY015390		INCYTE 015390
5	15193	47		HSFIB1		Fibronectin
6	15220	47		RRRPL9	R	Riboptn L9
7	15280	47		NCY015280		INCYTE 015280
8	15583	33		M62060		EST HHCH09 (IGR)
9	15662	31		HSACTCGR		Actin, gamma
10	15026	29		NCY015026		INCYTE 015026
11	15279	24		HSEF1AR		Elf 1-alpha
12	15027	23		NCY015027		INCYTE 015027
13	15033	20		NCY015033		INCYTE 015033
14	15198	20		NCY015198		INCYTE 015198
15	15809	20		HSCOLL1		Collagenase
16	15221	19		NCY015221		INCYTE 015221
17	15263	19		NCY015263		INCYTE 015263
18	15290	19		NCY015290		INCYTE 015290
19	15350	18		NCY015350		INCYTE 015350
20	15030	17		NCY015030		INCYTE 015030
21	15234	17		NCY015234		INCYTE 015234
22	15459	16		NCY015459		INCYTE 015459
23	15353	15		NCY015353		INCYTE 015353
24	15378	15		S76965		Ptn kinase inhib
25	15255	14		HUMTHYB4		Thymosin beta-4
26	15401	14		HSLIPCR		Lipocortin I
27	15425	14		HSPOLYAB		Poly-A bp
28	18212	14		HUMTHYMA		Thymosin, alpha
29	18216	14		HSMRP1		Motility relat ptn; MRP-1;CD-9
30	15189	13		HS18D		Interferon induc ptn 1-8D
31	15031	12		HUMFKBP		FK506 bp
32	15306	12		HSH2AZ		Histone H2A
33	15621	12		HUMLEC		Lectin, B-galbp, 14kDa
34	15789	11		NCY015789		INCYTE 015789
35	16578	11		HSRPS11		Riboptn S11
36	16632	11		M61984		EST HHCA13 (IGR)
37	18314	11		NCY018314		INCYTE 018314
38	15367	10		NCY015367		INCYTE 015367
39	15415	10		HSIFNIN1		interferon induc mRNA
40	15633	10		HSLDHAR		Lactate dehydrogenase
41	15813	10		CHKNMHCB		C Myosin heavy chain B
42	18210	10		NCY018210		INCYTE 018210
43	18233	10		HSRPII140		RNA polymerase II
44	18996	10		NCY018996		INCYTE 018996
45	15088	9		HUMFERL		Ferritin, light chain
46	15714	9		NCY015714		INCYTE 015714
47	15720	9		NCY015720		INCYTE 015720
48	15863	9		NCY015863		INCYTE 015863
49	16121	9		HSET		Endothelin
50	18252	9		NCY018252		INCYTE 018252
51	15351	8		HUMALBP		Lipid bp, adipocyte
52	15370	8		NCY015370		INCYTE 015370

TABLE 2 Con't

	number	N	c	entry	s	descriptor
53	15670	8		BTCIASHI	V	NADH-ubiq oxidoreductase
54	15795	8		NCY015795		INCYTE 015795
55	16245	8		NCY016245		INCYTE 016245
56	18262	8		NCY018262		INCYTE 018262
57	18321	8		HSRPL17		Riboptn L17
58	15126	7		XLRPL1BRF		Riboptn L1
59	15133	7		HSAC07		Actin, beta
60	15245	7		NCY015245		INCYTE 015245
61	15288	7		NCY015288		INCYTE 015288
62	15294	7		HSGAPDR		G-3-PD
63	15442	7		HUMLAMB		Laminin receptor, 54kDa
64	15485	7		HSNGMRNA		Uracil DNA glycosylase
65	16646	7		NCY016646		INCYTE 016646
66	18003	7		HUMPAIA		Plsmnogen activ gene
67	15032	6		HUMUB		Ubiquitin
68	15267	6		HSRPS8		Riboptn S8
69	15295	6		NCY015295		INCYTE 015295
70	15458	6		RNRPS10R	R	Riboptn S10
71	15832	6		RSGALEM	R	UDP-galactose epimerase
72	15928	6		HUMAPOJ		Apolipoptn J
73	16598	6		HUMTBMM40		Tubulin, beta
74	18218	6		NCY018218		INCYTE 018218
75	18499	6		HSP27		Hydrophobic ptn p27
76	18963	6		NCY018963		INCYTE 018963
77	18997	6		NCY018997		INCYTE 018997
78	15432	5		HSAGALAR		Galactosidase A, alpha
79	15475	5		NCY015475		INCYTE 015475
80	15721	5		NCY015721		INCYTE 015721
81	15865	5		NCY015865		INCYTE 015865
82	16270	5		NCY016270		INCYTE 016270
83	16886	5		NCY016886		INCYTE 016886
84	18500	5		NCY018500		INCYTE 018500
85	18503	5		NCY018503		INCYTE 018503
86	19672	5		RRRPL34	R	Riboptn L34
87	15086	4		XLRPL1AR	F	Riboptn L1a
88	15113	4		HUMIFNWRS		tRNA synthetase, trp
89	15242	4		NCY015242		INCYTE 015242
90	15249	4		NCY015249		INCYTE 015249
91	15377	4		NCY015377		INCYTE 015377
92	15407	4		NCY015407		INCYTE 015407
93	15473	4		NCY015473		INCYTE 015473
94	15588	4		HSRPS12		Riboptn S12
95	15684	4		HSEF1G		Elf 1-gamma
96	15782	4		NCY015782		INCYTE 015782
97	15916	4		HSRPS18		Riboptn S18
98	15930	4		NCY015930		INCYTE 015930
99	16108	4		NCY016108		INCYTE 016108
100	16133	4		NCY016133		INCYTE 016133

NORMAL MONONCYTE VS. ACTIVATED MACROPHAGE

Top 15 Most Abundant Genes

NORMAL

- 1 Elongation factor-I alpha
- 2 Ribosomal phosphoprotein
- 3 Ribosomal protein S8 homolog
- 4 Beta-Globin
- 5 Ferritin H chain
- 6 Ribosomal protein L7
- 7 Nucleoplasmin
- 8 Ribosomal protein S20 homolog
- 9 Transferrin receptor
- 10 Poly-A binding protein
- 11 Translationally controlled tumor ptn
- 12 Ribosomal protein S25
- 13 Signal recognition particle SRP9
- 14 Histone H2A.Z
- 15 Ribosomal protein Ke-3

ACTIVATED

- Interleukin-I beta
- Macrophage inflammatory protein-I
- Interleukin-8
- Lymphocyte activation gene
- Elongation factor-I alpha
- Beta actin
- Rantes T-cell specific protein
- Poly A binding protein
- Osteopontin; nephropontin
- Tumor Necrosis Factor-alpha
- INCYTE clone 011050
- Cu/Zn superoxide dismutase
- Adenylate cyclase (yeast homolog)
- NGF-related B cell activation molecule
- Protease Nexin-I, glial-derived

TABLE 3

TABLE 4

Libraries: THP-1
 Subtracting: HMC
 Sorted by ABUNDANCE
 Total clones analyzed: 7375

1057 genes, for a total of 2151 clones

number	entry	s descriptor	bgfreq	rfend	ratio
10022	HUMIL1	IL 1-beta	0	131	262.00
10036	HSMDNCF	IL-8	0	119	238.00
10089	HSLAG1CDN	Lymphocyte activ gene	0	71	142.00
10060	HUMTCSM	RANTES	0	23	46.000
10003	HUMMIPIA	MIP-1	3	121	40.333
10689	HSOP	Osteopontin	0	20	40.000
11050	NCY011050	INCYTE 011050	0	17	34.000
10937	HSTNFR	TNF-alpha	0	17	34.000
10176	HSSOD	Superoxide dismutase	0	14	28.000
10886	HSCDW40	B-cell activ,NGF-relat	0	10	20.000
10186	HUMAPR	Early resp PMA-induc	0	9	18.000
10967	HUMGDN	PN-1, glial-deriv	0	9	18.000
11353	NCY011353	INCYTE 011353	0	8	16.000
10298	NCY010298	INCYTE 010298	0	7	14.000
10215	HUM4COLA	Collagenase, type IV	0	6	12.000
10276	NCY010276	INCYTE 010276	0	6	12.000
10488	NCY010488	INCYTE 010488	0	6	12.000
11138	NCY011138	INCYTE 011138	0	6	12.000
10037	HUMCAPPRO	Adenylate cyclase	1	10	10.000
10840	HUMADCY	Adenylate cyclase	0	5	10.000
10672	HSCD44E	Cell adhesion glptn	0	5	10.000
12837	HUMCYCLOX	Cyclooxygenase-2	0	5	10.000
10001	NCY010001	INCYTE 010001	0	5	10.000
10005	NCY010005	INCYTE 010005	0	5	10.000
10294	NCY010294	INCYTE 010294	0	5	10.000
10297	NCY010297	INCYTE 010297	0	5	10.000
10403	NCY010403	INCYTE 010403	0	5	10.000
10699	NCY010699	INCYTE 010699	0	5	10.000
10966	NCY010966	INCYTE 010966	0	5	10.000
12092	NCY012092	INCYTE 012092	0	5	10.000
12549	HSRHOB	Oncogene rho	0	5	10.000
10691	HUMARF1BA	ADP-ribosylation fctr	0	4	8.000
12106	HSADSS	Adenylosuccinate synthetase	0	4	8.000
10194	HSCATHL	Cathepsin L	0	4	8.000
10479	CLMCYCA	I Cyclin A	0	4	8.000
10031	NCY010031	INCYTE 010031	0	4	8.000
10203	NCY010203	INCYTE 010203	0	4	8.000
10288	NCY010288	INCYTE 010288	0	4	8.000
10372	NCY010372	INCYTE 010372	0	4	8.000
10471	NCY010471	INCYTE 010471	0	4	8.000
10484	NCY010484	INCYTE 010484	0	4	8.000
10859	NCY010859	INCYTE 010859	0	4	8.000
10890	NCY010890	INCYTE 010890	0	4	8.000
11511	NCY011511	INCYTE 011511	0	4	8.000
11868	NCY011868	INCYTE 011868	0	4	8.000
12820	NCY012820	INCYTE 012820	0	4	8.000
10133	HSI1RAP	IL-1 antagonist	0	4	8.000
10516	HUMP2A	Phosphatase, regul 2A	0	4	8.000
11063	HUMB94	TNF-induc response	0	4	8.000
11140	HSHB15RNA	HB15 gene; new Ig	0	3	6.000
10788	NCY001713	INCYTE 001713	0	3	6.000
10033	NCY010033	INCYTE 010033	0	3	6.000
10035	NCY010035	INCYTE 010035	0	3	6.000
10084	NCY010084	INCYTE 010084	0	3	6.000
10236	NCY010236	INCYTE 010236	0	3	6.000
10383	NCY010383	INCYTE 010383	0	3	6.000

TABLE 4 Con't

number	entry	s descriptor	bqfreq	rfend	ratio
10450	NCY010450	INCYTE 010450	0	3	6.000
10470	NCY010470	INCYTE 010470	0	3	6.000
10504	NCY010504	INCYTE 010504	0	3	6.000
10507	NCY010507	INCYTE 010507	0	3	6.000
10598	NCY010598	INCYTE 010598	0	3	6.000
10779	NCY010779	INCYTE 010779	0	3	6.000
10909	NCY010909	INCYTE 010909	0	3	6.000
10976	NCY010976	INCYTE 010976	0	3	6.000
10985	NCY010985	INCYTE 010985	0	3	6.000
11052	NCY011052	INCYTE 011052	0	3	6.000
11068	NCY011068	INCYTE 011068	0	3	6.000
11134	NCY011134	INCYTE 011134	0	3	6.000
11136	NCY011136	INCYTE 011136	0	3	6.000
11191	NCY011191	INCYTE 011191	0	3	6.000
11219	NCY011219	INCYTE 011219	0	3	6.000
11386	NCY011386	INCYTE 011386	0	3	6.000
11403	NCY011403	INCYTE 011403	0	3	6.000
11460	NCY011460	INCYTE 011460	0	3	6.000
11618	NCY011618	INCYTE 011618	0	3	6.000
11686	NCY011686	INCYTE 011686	0	3	6.000
12021	NCY012021	INCYTE 012021	0	3	6.000
12025	NCY012025	INCYTE 012025	0	3	6.000
12320	NCY012320	INCYTE 012320	0	3	6.000
12330	NCY012330	INCYTE 012330	0	3	6.000
12853	NCY012853	INCYTE 012853	0	3	6.000
14386	NCY014386	INCYTE 014386	0	3	6.000
14391	NCY014391	INCYTE 014391	0	3	6.000

TABLE 5

```

* Master menu for SUBTRACTION output
SET TALK OFF
SET SAFETY OFF
SET EXACT ON
SET TYPEHEAD TO 0
CLEAR
SET DEVICE TO SCREEN
USE "SmartGuy:FoxBASE+/Mac:fox files:Clones.dbf"
GO TOP
STORE NUMBER TO INITIATE
GO BOTTOM
STORE NUMBER TO TERMINATE
STORE ' ' TO Target1
STORE ' ' TO Target2
STORE ' ' TO Target3
STORE ' ' TO Object1
STORE ' ' TO Object2
STORE ' ' TO Object3
STORE 0 TO ANAL
STORE 0 TO EMATCH
STORE 0 TO HMATCH
STORE 0 TO OMATCH
STORE 0 TO IMATCH
STORE 0 TO PTF
STORE 1 TO BAIL
DO WHILE .T.
  * Program.: Subtraction 2.fmt
  * Date.....: 10/11/94
  * Version.: FoxBASE+/Mac, revision 1.10
  * Notes.....: Format file Subtraction 2
  *
  SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Geneva",9 COLOR 0,0,0,
  @ PIXELS 75,120 TO 178,241 STYLE 3871 COLOR 0,0,-1,24610,-1,8947
  @ PIXELS 27,134 SAY "Subtraction Menu" STYLE 65536 FONT "Geneva",274 COLOR 0,0,-1,-1,-1,-1
  @ PIXELS 117,126 GET EMATCH STYLE 65536 FONT "Chicago",12 PICTURE "e*c Exact " SIZE 15,62 CO
  @ PIXELS 135,126 GET HMATCH STYLE 65536 FONT "Chicago",12 PICTURE "e*c Homologous" SIZE 15,1
  @ PIXELS 153,126 GET OMATCH STYLE 65536 FONT "Chicago",12 PICTURE "e*c Other spc" SIZE 15,84
  @ PIXELS 90,152 SAY "Matches:" STYLE 65536 FONT "Geneva",12 COLOR 0,0,-1,-1,-1,-1
  @ PIXELS 171,126 GET Imatch STYLE 65536 FONT "Chicago",12 PICTURE "e*c Incyte" SIZE 15,65 CO
  @ PIXELS 252,137 GET initiate STYLE 0 FONT "Geneva",12 SIZE 15,70 COLOR 0,0,-1,-1,-1,-1
  @ PIXELS 252,236 GET terminate STYLE 0 FONT "Geneva",12 SIZE 15,70 COLOR 0,0,-1,-1,-1,-1
  @ PIXELS 252,35 SAY "Include clones:" STYLE 65536 FONT "Geneva",12 COLOR 0,0,-1,-1,-1,-1
  @ PIXELS 252,215 SAY "->" STYLE 65536 FONT "Geneva",14 COLOR 0,0,-1,-1,-1,-1
  @ PIXELS 198,126 GET PTF STYLE 65536 FONT "Chicago",12 PICTURE "e*c Print to file" SIZE 15,9
  @ PIXELS 90,9 TO 181,109 STYLE 3871 COLOR 0,0,-1,-25600,-1,-1
  @ PIXELS 90,288 TO 181,397 STYLE 3871 COLOR 0,0,-1,-25600,-1,-1
  @ PIXELS 81,296 SAY "Background:" STYLE 65536 FONT "Geneva",270 COLOR 0,0,-1,-1,-1,-1
  @ PIXELS 45,135 GET ANAL STYLE 65536 FONT "Chicago",12 PICTURE "e*R Overall;Function" SIZE 4
  @ PIXELS 81,26 SAY "Target:" STYLE 65536 FONT "Geneva",270 COLOR 0,0,-1,-1,-1,-1
  @ PIXELS 108,20 GET target1 STYLE 0 FONT "Geneva",9 SIZE 12,79 COLOR 0,0,-1,-1,-1,-1
  @ PIXELS 135,20 GET target2 STYLE 0 FONT "Geneva",9 SIZE 12,79 COLOR 0,0,-1,-1,-1,-1
  @ PIXELS 162,20 GET target3 STYLE 0 FONT "Geneva",9 SIZE 12,79 COLOR 0,0,-1,-1,-1,-1
  @ PIXELS 108,299 GET object1 STYLE 0 FONT "Geneva",9 SIZE 12,79 COLOR 0,0,-1,-1,-1,-1
  @ PIXELS 135,299 GET object2 STYLE 0 FONT "Geneva",9 SIZE 12,79 COLOR 0,0,-1,-1,-1,-1
  @ PIXELS 162,299 GET object3 STYLE 0 FONT "Geneva",9 SIZE 12,79 COLOR 0,0,-1,-1,-1,-1
  @ PIXELS 276,324 GET Bail STYLE 65536 FONT "Chicago",12 PICTURE "e*R Run;Bail out" SIZE 4112
  *
  * EOF: Subtraction 2.fmt
  READ
  IF Bail=2
    CLEAR
    CLOSE DATABASES
    USE "SmartGuy:FoxBASE+/Mac:fox files:clones.dbf"
    SET SAFETY ON
    SCREEN 1 OFF
    RETURN

```



```

ENDIF
STORE VAL(SYS(2)) TO STARTIME
STORE UPPER(Target1) TO Target1
STORE UPPER(Target2) TO Target2
STORE UPPER(Target3) TO Target3
STORE UPPER(Object1) TO Object1
STORE UPPER(Object2) TO Object2
STORE UPPER(Object3) TO Object3
clear
SET TALK ON
GAP = TERMINATE-INITIATE+1
GO INITIATE
COPY NEXT GAP FIELDS NUMBER,library,D,F,Z,R,ENTRY,S,DESCRIPTOR,START,RFEND,I TO TEMPNUM
USE TEMPNUM
COUNT TO TOT
COPY TO TEMPRED FOR D='E'.OR.D='O'.OR.D='H'.OR.D='N'.OR.D='I'
USE TEMPRED

IF Bmatch=0 .AND. Hmatch=0 .AND. Qmatch=0 .AND. IMATCH=0
COPY TO TEMPDESIG
ELSE
COPY STRUCTURE TO TEMPDESIG
USE TEMPDESIG
  IF Bmatch=1
    APPEND FROM TEMPNUM FOR D='E'
  ENDIF
  IF Hmatch=1
    APPEND FROM TEMPNUM FOR D='H'
  ENDIF
  IF Qmatch=1
    APPEND FROM TEMPNUM FOR D='O'
  ENDIF
  IF Imatch=1
    APPEND FROM TEMPNUM FOR D='I'.OR.D='X'
    *.OR.D='N'
  ENDIF
ENDIF
COUNT TO STARTTOT

COPY STRUCTURE TO TEMPLIB
USE TEMPLIB
  APPEND FROM TEMPDESIG FOR library=UPPER(target1)
  IF target2<>'
    APPEND FROM TEMPDESIG FOR library=UPPER(target2)
  ENDIF
  IF target3<>'
    APPEND FROM TEMPDESIG FOR library=UPPER(target3)
  ENDIF
COUNT TO ANALTOT

USE TEMPDESIG
COPY STRUCTURE TO TEMPSUB
USE TEMPSUB
  APPEND FROM TEMPDESIG FOR library=UPPER(Object1)
  IF target2<>'
    APPEND FROM TEMPDESIG FOR library=UPPER(Object2)
  ENDIF
  IF target3<>'
    APPEND FROM TEMPDESIG FOR library=UPPER(Object3)
  ENDIF
COUNT TO SUBTRACTOT
SET TALK OFF
*****
* COMPRESSION SUBROUTINE A
? 'COMPRESSING QUERY LIBRARY'
USE TEMPLIB

```

```

SORT ON ENTRY,NUMBER TO LIBSORT
USE LIBSORT
COUNT TO IDGENE
REPLACE ALL RFEND WITH 1
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
  IF MARK1 >= IDGENE
    PACK
    COUNT TO AUNIQUE
    SW2=1
  LOOP
  ENDIF
GO MARK1
DUP = 1
STORE ENTRY TO TESTA
STORE D TO DESIGA
SW = 0
DO WHILE SW=0 TEST
  SKIP
  STORE ENTRY TO TESTB
  STORE D TO DESIGB
  IF TESTA = TESTB.AND.DESIGA=DESIGB
    DELETE
    DUP = DUP+1
  LOOP
  ENDIF
GO MARK1
REPLACE RFEND WITH DUP
MARK1 = MARK1+DUP
SW=1
LOOP
ENDDO TEST
LOOP
ENDDO ROLL
SORT ON RFEND/D,NUMBER TO TEMPTARSORT
USE TEMPTARSORT
*REPLACE ALL START WITH RFEND/IDGENE*10000
COUNT TO TEMPTARCO
*****
* COMPRESSION SUBROUTINE B
? 'COMPRESSING TARGET LIBRARY'
USE TEMPSUB
SORT ON ENTRY,NUMBER TO SUBSORT
USE SUBSORT
COUNT TO SUBGENE
REPLACE ALL RFEND WITH 1
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
  IF MARK1 >= SUBGENE
    PACK
    COUNT TO BUNIQUE
    SW2=1
  LOOP
  ENDIF
GO MARK1
DUP = 1
STORE ENTRY TO TESTA
STORE D TO DESIGA
SW = 0
DO WHILE SW=0 TEST
  SKIP
  STORE ENTRY TO TESTB
  STORE D TO DESIGB
  IF TESTA = TESTB.AND.DESIGA=DESIGB

```

```

DELETE
DUP = DUP+1
LOOP
ENDIF
GO MARK1
REPLACE RFEND WITH DUP
MARK1 = MARK1+DUP
SW=1
LOOP
ENDDO TEST
LOOP :
ENDDO ROLL
SORT ON RFEND/D,NUMBER TO TEMPSUBSORT
USE TEMPSUBSORT
*REPLACE ALL START WITH RFEND/IDGENE*10000
COUNT TO TEMPSUBCO
*****
*FUSION ROUTINE
? 'SUBTRACTING LIBRARIES'
USE SUBTRACTION
COPY STRUCTURE TO CRUNCHER
SELECT 2
USE TEMPSUBSORT
SELECT 1:
USE CRUNCHER
APPEND FROM TEMPTARSORT
COUNT TO BAILOUT
MARK = 0

DO WHILE .T.
SELECT 1
MARK = MARK+1
IF MARK>BAILOUT
EXIT
ENDIF
GO MARK
STORE ENTRY TO SCANNER
SELECT 2
LOCATE FOR ENTRY=SCANNER
IF FOUND()
STORE RFEND TO BIT1
STORE RFEND TO BIT2
ELSE
STORE 1/2 TO BIT1
STORE 0 TO BIT2
ENDIF
SELECT 1
REPLACE BGFREQ WITH BIT2
REPLACE ACTUAL WITH BIT1
LOOP
ENDDO

SELECT 1
REPLACE ALL RATIO WITH RFEND/ACTUAL
? 'DOING FINAL SORT BY RATIO'
SORT ON RATIO/D,BGFREQ/D,DESCRIPTOR TO FINAL
USE FINAL
*****
set talk off
DO CASE
CASE PTF=0
SET DEVICE TO PRINT
SET PRINT ON
EJECT
CASE PTF=1
SET ALTERNATE TO "Adenoid.Patent.Figures:Subtraction.txt"

```

```
SET ALTERNATE ON
ENDCASE
```

```
STORE VAL(SYS(2)) TO FINTIME
IF FINTIME<STARTIME
STORE FINTIME+86400 TO FINTIME
ENDIF
STORE FINTIME - STARTIME TO COMPSEC
STORE COMPSEC/60 TO COMPMIN
```

```
*****
```

```
SET MARGIN TO 10
```

```
61,1 SAY "Library Subtraction Analysis" STYLE 65536 FONT "Geneva",274 COLOR 0,0,0,-1,-1,-1
```

```
?
```

```
?
```

```
?
```

```
?
```

```
?
```

```
??
```

```
?? TIME()
```

```
?
```

```
?? STR(INITIATE,5,0)
```

```
??
```

```
?? STR(TERMINATE,6,0)
```

```
?
```

```
?
```

```
IF Target2<>
```

```
??
```

```
?? Target2
```

```
ENDIF
```

```
IF Target3<>
```

```
??
```

```
?? Target3
```

```
ENDIF
```

```
?
```

```
?
```

```
IF Object2<>
```

```
??
```

```
?? Object2
```

```
ENDIF
```

```
IF Object3<>
```

```
??
```

```
?? Object3
```

```
ENDIF
```

```
?
```

```
IF Ematch=0 .AND. Hmatch=0 .AND. Omatch=0 .AND. IMATCH=0
```

```
??
```

```
IF Ematch=1
```

```
??
```

```
ENDIF
```

```
IF Hmatch=1
```

```
??
```

```
ENDIF
```

```
IF Omatch=1
```

```
??
```

```
ENDIF
```

```
IF Imatch=1
```

```
??
```

```
ENDIF
```

```
IF ANAL=1
```

```
?
```

```
ENDIF
```

```
IF ANAL=2
```

```
?
```

```
ENDIF
```

53

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Helvetica",265 COLOR 0
 ? 'Viral elements:'
 SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Geneva",7 COLOR 0,0,0,
 list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R="V" Page 7 of 38

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Helvetica",265 COLOR 0
 ? 'Kinases and Phosphatases:'
 SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Geneva",7 COLOR 0,0,0,
 list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R="Y"

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Helvetica",265 COLOR 0
 ? 'Tumor-related antigens:'
 SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Geneva",7 COLOR 0,0,0,
 list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R="A"

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Helvetica",268 COLOR 0
 ?
 ? PROTEIN SYNTHETIC MACHINERY PROTEINS'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Helvetica",265 COLOR 0
 ? 'Transcription and Nucleic Acid-binding proteins:'
 SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Geneva",7 COLOR 0,0,0,
 list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R="D"

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Helvetica",265 COLOR 0
 ? 'Translation:'
 SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Geneva",7 COLOR 0,0,0,
 list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R="T"

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Helvetica",265 COLOR 0
 ? 'Ribosomal proteins:'
 SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Geneva",7 COLOR 0,0,0,
 list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R="R"

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Helvetica",265 COLOR 0
 ? 'Protein processing:'
 SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Geneva",7 COLOR 0,0,0,
 list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R="L"

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Helvetica",268 COLOR 0
 ?
 ? ENZYMES'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Helvetica",265 COLOR 0
 ? 'Ferropoteins:'
 SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Geneva",7 COLOR 0,0,0,
 list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R="F"

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Helvetica",265 COLOR 0
 ? 'Proteases and inhibitors:'
 SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Geneva",7 COLOR 0,0,0,
 list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R="P"

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Helvetica",265 COLOR 0
 ? 'Oxidative phosphorylation:'
 SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Geneva",7 COLOR 0,0,0,
 list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R="Z"

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Helvetica",265 COLOR 0
 ? 'Sugar metabolism:'
 SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Geneva",7 COLOR 0,0,0,
 list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R="Q"

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Helvetica",265 COLOR 0
 ? 'Amino acid metabolism:'
 SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Geneva",7 COLOR 0,0,0,

list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='M'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Helvetica",265 COLOR 0
 ? "Nucleic acid metabolism;"

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Geneva",7 COLOR 0,0,0,
 list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='N'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Helvetica",265 COLOR 0
 ? "Lipid metabolism;"

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Geneva",7 COLOR 0,0,0,
 list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='W'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Helvetica",265 COLOR 0
 ? "Other enzymes;"

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Geneva",7 COLOR 0,0,0,
 list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='E'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Helvetica",268 COLOR 0
 ?
 ?
 ? MISCELLANEOUS CATEGORIES'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Helvetica",265 COLOR 0
 ? "Stress response;"

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Geneva",7 COLOR 0,0,0,
 list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='H'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Helvetica",265 COLOR 0
 ? "Structural;"

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Geneva",7 COLOR 0,0,0,
 list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='K'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Helvetica",265 COLOR 0
 ? "Other clones;"

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Geneva",7 COLOR 0,0,0,
 list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='X'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Helvetica",265 COLOR 0
 ? "Clones of unknown function;"

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Geneva",7 COLOR 0,0,0,
 list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='U'

ENDCASE

DO "Test print.prp"
 SET PRINT OFF
 SET DEVICE TO SCREEN
 CLOSE DATABASES
 ERASE TEMPLIB.DBF
 ERASE TEMPNUM.DBF
 ERASE TEMPEDESIG.DBF
 SET MARGIN TO 0
 CLEAR
 LOOP
 ENDDO

```

*Northern (single), version 11-25-94
close databases
SET TALK OFF
SET PRINT OFF
SET EXACT OFF
CLEAR
STORE '          ' TO Eobject
STORE '          ' TO Dobject
STORE 0 TO Numb.
STORE 0 TO Zog
STORE 1 TO Bail
DO WHILE .T.
* Program.: Northern (single).fmt
* Date....: 8/ 8/94
* Version.: FoxBASE+/Mac, revision 1.10
* Notes....: Format file Northern (single)
*
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Geneva",12 COLOR 0,0,0
@ PIXELS 15,81 TO 46,397 STYLE 28447 COLOR 0,0,-1,-25600,-1,-1
@ PIXELS 89,79 TO 192,422 STYLE 28447 COLOR 0,0,0,-25600,-1,-1
@ PIXELS 115,98 SAY "Entry #:" STYLE 65536 FONT "Geneva",12 COLOR 0,0,0,-1,-1,-1
@ PIXELS 115,173 GET Eobject STYLE 0 FONT "Geneva",12 SIZE 15,142 COLOR 0,0,0,-1,-1,-1
@ PIXELS 145,89 SAY "Description" STYLE 65536 FONT "Geneva",12 COLOR 0,0,0,-1,-1,-1
@ PIXELS 145,173 GET Dobject STYLE 0 FONT "Geneva",12 SIZE 15,241 COLOR 0,0,0,-1,-1,-1
@ PIXELS 35,89 SAY "Single Northern search screen" STYLE 65536 FONT "Geneva",274 COLOR 0,0,-
@ PIXELS 220,162 GET Bail STYLE 65536 FONT "Chicago",12 PICTURE "@*R Continue;Bail out" SIZE
@ PIXELS 175,98 SAY "Clone #:" STYLE 65536 FONT "Geneva",12 COLOR 0,0,0,-1,-1,-1
@ PIXELS 175,173 GET Numb STYLE 0 FONT "Geneva",12 SIZE 15,70 COLOR 0,0,0,-1,-1,-1
@ PIXELS 80,152 SAY "Enter any ONE of the following:" STYLE 65536 FONT "Geneva",12 COLOR -1,
*
* EOF: Northern (single).fmt
READ
IF Bail=2
CLEAR
screen 1 off
RETURN
ENDIF
USE "SmartGuy\FoxBASE+/Mac\Fox files\Lookup.dbf"
SET TALK ON

IF Eobject<>'
STORE UPPER(Eobject) to Eobject
SET SAFETY OFF
SORT ON Entry TO "Lookup entry.dbf"
SET SAFETY ON
USE "Lookup entry.dbf"
LOCATE FOR Look=Eobject
IF .NOT.FOUND()
CLEAR
LOOP
ENDIF
BROWSE
STORE Entry TO Searchval.
CLOSE DATABASES
ERASE "Lookup entry.dbf"
ENDIF

IF Dobject<>'
SET EXACT OFF
SET SAFETY OFF
SORT ON descriptor TO "Lookup descriptor.dbf"
SET SAFETY On
USE "Lookup descriptor.dbf"
LOCATE FOR UPPER(TRIM(descriptor))=UPPER(TRIM(Dobject))
IF .NOT.FOUND()
CLEAR

```



```

LOOP
ENDIF
BROWSE
STORE Entry TO Searchval
CLOSE DATABASES
ERASE 'Lookup descriptor.dbf'
SET EXACT ON
ENDIF

IF Numb<=0
USE 'SmartGuy\FoxBASE+Mac\Fox files:clones.dbf'
GO Numb
BROWSE
STORE Entry TO Searchval
ENDIF

CLEAR
? 'Northern analysis for entry '
?? Searchval
?
? 'Enter Y to proceed'
WAIT TO OK
CLEAR
IF UPPER(OK) <= 'Y'
screen 1 off
RETURN
ENDIF

* COMPRESSION SUBROUTINE FOR Library.dbf
? 'Compressing the Libraries file now...'
USE 'SmartGuy\FoxBASE+Mac\Fox files:libraries.dbf'
SET SAFETY OFF
SORT ON library TO 'Compressed libraries.dbf'
* FOR entered=0
SET SAFETY ON
USE 'Compressed libraries.dbf'
DELETE FOR entered=0
PACK
COUNT TO TOT
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
  IF MARK1 >= TOT
    PACK
    SW2=1
  LOOP
  ENDO
GO MARK1
STORE library TO TESTA
SKIP
STORE Library TO TESTB
IF TESTA = TESTB
DELETE
ENDIF
MARK1 = MARK1+1
LOOP
ENDDO ROLL

* Northern analysis
CLEAR
? 'Doing the northern now...'
SET TALK ON
USE 'SmartGuy\FoxBASE+Mac\Fox files:clones.dbf'
SET SAFETY OFF
COPY TO 'Hits.dbf' FOR entry=searchval
SET SAFETY ON

```

```

* MASTER ANALYSIS 3; VERSION 12-9-94
* Master menu for analysis output
CLOSE DATABASES
SET TALK OFF
SET SAFETY OFF
CLEAR
SET DEVICE TO SCREEN
SET DEFAULT TO "SmartGuy:FoxBASE+/Mac:fox files:Output programs;"
USE "SmartGuy:FoxBASE+/Mac:fox files:Clones.dbf"
GO TOP
STORE NUMBER TO INITIATE
GO BOTTOM
STORE NUMBER TO TERMINATE
STORE 0 TO ENTIRE
STORE 0 TO CONDEN
STORE 0 TO ANAL
STORE 0 TO EMATCH
STORE 0 TO HMATCH
STORE 0 TO OMATCH
STORE 0 TO IMATCH
STORE 0 TO XMATCH
STORE 0 TO PRINTON
STORE 0 TO PTF
DO WHILE .T.
* Program.: Master analysis.fmt
* Date.....: 12/ 9/94
* Version.: FoxBASE+/Mac, revision 1.10
* Notes.....: Format file Master analysis
*
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Geneva",9 COLOR 0,0,0,
@ PIXELS 39,255 TO 277,430 STYLE 28447 COLOR 0,0,-1,-25600,-1,-1
@ PIXELS 75,120 TO 178,241 STYLE 3871 COLOR 0,0,-1,-25600,-1,-1
@ PIXELS 27,98 SAY "Customized Output Menu" STYLE 65536 FONT "Geneva",274 COLOR 0,0,-1,-1,-1
@ PIXELS 45,54 GET conden STYLE 65536 FONT "Chicago",12 PICTURE "@*C Condensed format" SIZE
@ PIXELS 54,261 GET anal STYLE 65536 FONT "Chicago",12 PICTURE "@*RV Sort/number:Sort/entry;"
@ PIXELS 117,126 GET EMATCH STYLE 65536 FONT "Chicago",12 PICTURE "@*C Exact " SIZE 15,62 CO
@ PIXELS 135,126 GET HMATCH STYLE 65536 FONT "Chicago",12 PICTURE "@*C Homologous" SIZE 15,1
@ PIXELS 153,126 GET OMATCH STYLE 65536 FONT "Chicago",12 PICTURE "@*C Other spc" SIZE 15,84
@ PIXELS 90,152 SAY "Matches:" STYLE 65536 FONT "Geneva",268 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 63,54 GET PRINTON STYLE 65536 FONT "Chicago",12 PICTURE "@*C Include clone listing"
@ PIXELS 171,126 GET Imatch STYLE 65536 FONT "Chicago",12 PICTURE "@*C Incyte" SIZE 15,65 CO
@ PIXELS 252,146 GET initiate STYLE 0 FONT "Geneva",12 SIZE 15,70 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 270,146 GET terminate STYLE 0 FONT "Geneva",12 SIZE 15,70 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 234,134 SAY "Include clones " STYLE 65536 FONT "Geneva",12 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 270,125 SAY "->" STYLE 65536 FONT "Geneva",14 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 198,126 GET PTF STYLE 65536 FONT "Chicago",12 PICTURE "@*C Print to file" SIZE 15,9
@ PIXELS 189,0 TO 257,120 STYLE 3871 COLOR 0,0,-1,-25600,-1,-1
@ PIXELS 209,8 SAY "Library selection" STYLE 65536 FONT "Geneva",266 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 227,18 GET ENTIRE STYLE 65536 FONT "Chicago",12 PICTURE "@*RV All;Selected" SIZE 16
*
* EOF: Master analysis.fmt
READ
IF ANAL=9
CLEAR
CLOSE DATABASES
ERASE TEMPMASTER.DBF
USE "SmartGuy:FoxBASE+/Mac:fox files:clones.dbf"
SET SAFETY ON
SCREEN 1 OFF
RETURN
ENDIF
clear
? INITIATE
? TERMINATE
? CONDEN
? ANAL

```

```

? ematch
? Hmatch
? Cmatch
? IMATCH
SET TALK ON
IF ENTIRE=2
USE "Unique libraries.dbf"
REPLACE ALL i WITH ' '
BROWSE FIELDS i, libname, library, total, entered AT 0,0
ENDIF
USE "SmartGuy:FoxBASE+/Mac:fox files:clones.dbf"
*COPY TO TEMPNUM FOR NUMBER>=INITIATE.AND.NUMBER<=TERMINATE
*USE TEMPNUM
COPY STRUCTURE TO TEMPLIB
USE TEMPLIB
IF ENTIRE=1
APPEND FROM "SmartGuy:FoxBASE+/Mac:fox files:Clones.dbf"
ENDIF
IF ENTIRE=2
USE "Unique libraries.dbf"
COPY TO SELECTED FOR UPPER(i)='Y'
USE SELECTED
STORE RECCOUNT() TO STOPIT
MARK=1
DO WHILE .T.
IF MARK>STOPIT
CLEAR
EXIT
ENDIF
USE SELECTED
GO MARK
STORE library TO THISONE
? 'COPYING '
?? THISONE
USE TEMPLIB
APPEND FROM "SmartGuy:FoxBASE+/Mac:fox files:Clones.dbf" FOR library=THISONE
STORE MARK+1 TO MARK
LOOP
ENDDO
ENDIF
USE "SmartGuy:FoxBASE+/Mac:fox files:clones.dbf"
COUNT TO STARTOT
COPY STRUCTURE TO TEMPDESIG
USE TEMPDESIG
IF Ematch=0 .AND. Hmatch=0 .AND. Cmatch=0 .AND. IMATCH=0
APPEND FROM TEMPLIB
ENDIF
IF Ematch=1
APPEND FROM TEMPLIB FOR D='E'
ENDIF
IF Hmatch=1
APPEND FROM TEMPLIB FOR D='H'
ENDIF
IF Cmatch=1
APPEND FROM TEMPLIB FOR D='O'
ENDIF
IF Imatch=1
APPEND FROM TEMPLIB FOR D='I'.OR.D='X'.OR.D='N'
ENDIF
IF Xmatch=1
APPEND FROM TEMPLIB FOR D='X'
ENDIF
COUNT TO ANALTOT
set talk off
*****
DO CASE

```

```

CASE PTF=0
SET DEVICE TO PRINT
SET PRINT ON
EJECT
CASE PTF=1
SET ALTERNATE TO "Total function sort.txt"
*SET ALTERNATE TO "H and O function sort.txt"
*SET ALTERNATE TO "Shear Stress HUVEC 2:Abundance sort.txt"
*SET ALTERNATE TO "Shear Stress HUVEC 2:Abundance con.txt"
*SET ALTERNATE TO "Shear Stress HUVEC 2:Function sort.txt"
*SET ALTERNATE TO "Shear Stress HUVEC 2:Distribution sort.txt"
*SET ALTERNATE TO "Shear stress HUVEC 1:Clone list.txt"
*SET ALTERNATE TO "Shear Stress HUVEC 2:Location sort.txt"
SET ALTERNATE ON
ENDCASE
*****
IF PRINTON=1
@1,30 SAY "Database Subset Analysis" STYLE 65536 FONT "Geneva",274 COLOR 0,0,0,-1,-1,-1
ENDIF
?
?
?
?
? date()
?? '
?? TIME()
? 'Clone numbers '
?? STR(INITIAL,6,0)
?? ' through '
?? STR(TERMINATE,6,0)
? 'Libraries: '
IF ENTIRE=1
? 'All libraries'
ENDIF
IF ENTIRE=2
MARK=1
DO WHILE .T.
IF MARK>STOPIT
EXIT
ENDIF
USE SELECTED
GO MARK
? '
?? TRIM(libname)
STORE MARK+1 TO MARK
LOOP
ENDDO
ENDIF
? 'Designations: '
IF Ematch=0 .AND. Hmatch=0 .AND. Omatch=0 .AND. IMATCH=0
?? 'All'
ENDIF
IF Ematch=1
?? 'Exact,'
ENDIF
IF Hmatch=1
?? 'Human,'
ENDIF
IF Omatch=1
?? 'Other sp.'
ENDIF
IF Imatch=1
?? 'INCYTE'
ENDIF
IF Xmatch=1
?? 'EST'

```

```

ENDIF
IF CONDEN=1
? 'Condensed format analysis'
ENDIF
IF ANAL=1
? 'Sorted by NUMBER'
ENDIF
IF ANAL=2
? 'Sorted by ENTRY'
ENDIF
IF ANAL=3
? 'Arranged by ABUNDANCE'
ENDIF
IF ANAL=4
? 'Sorted by INTEREST'
ENDIF
IF ANAL=5
? 'Arranged by LOCATION'
ENDIF
IF ANAL=6
? 'Arranged by DISTRIBUTION'
ENDIF
IF ANAL=7
? 'Arranged by FUNCTION'
ENDIF
? 'Total clones represented: '
?? STR(STARTOT,6,0)
? 'Total clones analyzed: '
?? STR(ANALTOT,6,0)
?
? '1 = library    d = designation    f = distribution    z = location    r = function    c = cer
?
*****
USE TEMPDDESIG
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Geneva",7 COLOR 0,0,0,
DO CASE
CASE ANAL=1
* sort/number
SET HEADING ON
IF CONDEN=1
SORT TO TEMP1 ON ENTRY,NUMBER
DO "COMPRESSION number.PRG"
ELSE
SORT TO TEMP1 ON NUMBER
USE TEMP1
list off fields number,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR
*list off fields number,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,RFEND,INIT,I
CLOSE DATABASES
ERASE TEMP1.DBF
ENDIF

CASE ANAL=2
* sort/DESCRIPTOR
SET HEADING ON
*SORT TO TEMP1 ON DESCRIPTOR,ENTRY,NUMBER/S for D='E'.OR.D='H'.OR.D='O'.OR.D='X'.OR.D='I'
*SORT TO TEMP1 ON ENTRY,DESCRIPTOR,NUMBER/S for D='E'.OR.D='H'.OR.D='O'.OR.D='X'.OR.D='I'
SORT TO TEMP1 ON ENTRY,START/S for D='E'.OR.D='H'.OR.D='O'.OR.D='X'.OR.D='I'
IF CONDEN=1
DO "COMPRESSION entry.PRG"
ELSE
USE TEMP1
list off fields number,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,RFEND,INIT,I
CLOSE DATABASES
ERASE TEMP1.DBF
ENDIF

```

```

CASE ANAL=3
* sort by abundance
SET HEADING ON
SORT TO TEMP1 ON ENTRY,NUMBER FOR D='E'.OR.D='H'.OR.D='O'.OR.D='X'.OR.D='I'
DO "COMPRESSION abundance.PRG"

CASE ANAL=4
* sort/interest
SET HEADING ON
IF CONDEN=1
SORT TO TEMP1 ON ENTRY,NUMBER FOR I>0
DO "COMPRESSION interest.PRG"
ELSE
SORT ON I/D,ENTRY TO TEMP1 FOR I>1
USE TEMP1
list off fields number,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,RFEND,INIT,I
CLOSE DATABASES
ERASE TEMP1.DBF
ENDIF

CASE ANAL=5
* arrange/location
SET HEADING ON
STORE 4 TO AMPLIFIER
? 'Nuclear:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression location.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Cytoplasmic:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression location.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Cytoskeleton:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression location.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Cell surface:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression location.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Intracellular membrane:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression location.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Mitochondrial:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression location.prg"
ELSE
DO "Normal subroutine 1"
ENDIF

```

```

? 'Secreted;'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression location.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Other;'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression location.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Unknown;'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression location.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
IF CONDEN=1
SET DEVICE TO PRINTER
SET PRINTER ON
EJECT
DO "Output heading.prg"
USE "Analysis location.dbf"
DO "Create bargraph.prg"
SET HEADING OFF
? '          FUNCTIONAL CLASS          TOTAL    UNIQUE    NEW    % TOTAL'
?
LIST OFF FIELDS Z,NAME,CLONES,GENES,NEW,PERCENT,GRAPH
CLOSE DATABASES
ERASE TEMP2.DBF
SET HEADING ON
*USE "SmartGuy:FoxBASE+/Mac:fox files:TEMPMASTER.dbf"
ENDIF

CASE ANAL=6
* arrange/distribution
SET HEADING ON
STORE 3 TO AMPLIFIER
? 'Cell/tissue specific distribution:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression distrib.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Non-specific distribution:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression distrib.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Unknown distribution:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression distrib.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
IF CONDEN=1
SET DEVICE TO PRINTER
SET PRINTER ON

```

```

EJECT
DO "Output heading.prg"
USE "Analysis distribution.dbf"
DO "Create bargraph.prg"
SET HEADING OFF
? '          FUNCTIONAL CLASS                      TOTAL    UNIQUE    $ TOTAL'
?
LIST OFF FIELDS P,NAME,CLONES,GENES,PERCENT,GRAPH
CLOSE DATABASES
ERASE TEMP2.DBF
SET HEADING ON
*USE "SmartGuy:FoxBASE+/Mac:fox files:TEMPMASTER.dbf"
ENDIF

CASE ANAL=7
* arrange/function
SET HEADING ON
STORE 10 TO AMPLIFIER
? '          BINDING PROTEINS'
?
? 'Surface molecules and receptors:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDENSE=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Calcium-binding proteins:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDENSE=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Ligands and effectors:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDENSE=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Other binding proteins:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDENSE=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
EJECT
? '          ONCOGENES'
?
? 'General oncogenes:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDENSE=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'GTP-binding proteins:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDENSE=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Viral elements:'

```



```

SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Kinases and Phosphatases:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Tumor-related antigens:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
*EJECT
? '
                                PROTEIN SYNTHETIC MACHINERY PROTEINS'
?
? 'Transcription and Nucleic Acid-binding proteins:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Translation:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Ribosomal proteins:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Protein processing:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
*EJECT
? '
                                ENZYMES'
?
? 'Ferropoteins:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Proteases and inhibitors:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE

```

```

DO "Normal subroutine 1"
ENDIF
? 'Oxidative phosphorylation:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Sugar metabolism:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Amino acid metabolism:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Nucleic acid metabolism:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Lipid metabolism:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Other enzymes:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
*EJECT
? '
?
? 'Stress response:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Structural:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Other clones:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE

```

MISCELLANEOUS CATEGORIES'

```

DO "Normal subroutine 1"
ENDIF
? "Clones of unknown function:"
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF

IF CONDEN=1
EJECT
*SET DEVICE TO PRINTER
*SET PRINT ON
DO "Output heading.prg"
***
USE "Analysis function.dbf"
DO "Create bargraph.prg"
SET HEADING OFF
***
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Geneva",12 COLOR 0,0,0
***
? '
? '          FUNCTIONAL CLASS
? '          CLONES      GENES      GENES      TOTAL      TOTAL      NEW      DIST
? '
***
*LIST OFF FIELDS P,NAME,CLONES,GENES,NEW,PERCENT,GRAPH,COMPANY
LIST OFF FIELDS P,NAME,CLONES,GENES,NEW,PERCENT,GRAPH
CLOSE DATABASES
ERASE TEMP2.DBF
SET HEADING ON
*USE "SmartGuy:FoxBASE+/Mac:fox files:TEMPMASTER.dbf"
ENDIF
CASE ANAL=8
DO "Subgroup summary 3.prg"
ENDCASE
DO "Test print.prg"
SET PRINT OFF
SET DEVICE TO SCREEN
CLOSE DATABASES
*ERASE TEMPLIB.DBF
*ERASE TEMPNUM.DBF
*ERASE TEMPDESIG.DBF
*ERASE SELECTED.DBF
CLEAR
LOOP
ENDDO

```

```

* COMPRESSION SUBROUTINE FOR ANALYSIS PROGRAMS
USE TEMP1
COUNT TO TOT
REPLACE ALL RFEND WITH 1
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
  IF MARK1 >= TOT
    PACK
    COUNT TO UNIQUE
    COUNT TO NEWGENES FOR D='H'.OR.D='O'
    SW2=1
    LOOP
  ENDIF
GO MARK1
DUP = 1
STORE ENTRY TO TESTA
SW = 0
DO WHILE SW=0 TEST
  SKIP
  STORE ENTRY TO TESTB
  IF TESTA = TESTB
    DELETE
    DUP = DUP+1
    LOOP
  ENDIF
GO MARK1.
REPLACE RFEND WITH DUP
MARK1 = MARK1+DUP
SW=1
LOOP
ENDDO TEST
LOOP
ENDDO ROLL
GO TOP
STORE Z TO LOC
USE "Analysis location.dbf"
LOCATE FOR Z=LOC
REPLACE CLONES WITH TOT
REPLACE GENES WITH UNIQUE
REPLACE NEW WITH NEWGENES
USE TEMP1
SORT ON RFEND/D TO TEMP2
USE TEMP2
?? STR(UNIQUE,5,0)
?? ' genes, for a total of '
?? STR(TOT,5,0)
?? ' clones'
? ' V Coincidence'
list off fields number,RFEND,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I

*SET PRINT OFF
CLOSE DATABASES
ERASE TEMP1.DBF
ERASE TEMP2.DBF
USE TEMPDESIG

```

* COMPRESSION SUBROUTINE FOR ANALYSIS PROGRAMS

USE TEMP1

COUNT TO TOT

REPLACE ALL RFEND WITH 1

MARK1 = 1

SW2=0

DO WHILE SW2=0 ROLL

IF MARK1 >= TOT

PACK

COUNT TO UNIQUE

SW2=1

LOOP

ENDIF

GO MARK1

DUP = 1

STORE ENTRY TO TESTA

SW = 0

DO WHILE SW=0 TEST

SKIP

STORE ENTRY TO TESTB

IF TESTA = TESTB

DELETE

DUP = DUP+1

LOOP

ENDIF

GO MARK1

REPLACE RFEND WITH DUP

MARK1 = MARK1+DUP

SW=1

LOOP

ENDDO TEST

LOOP

ENDDO ROLL

*BROWSE

*SET PRINTER ON

SORT ON DATE TO TEMP2

USE TEMP2

?? STR(UNIQUE,4,0)

?? ' genes, for a total of'

?? STR(TOT,4,0)

?? ' clones'

?

? ' V Coincidence'

COUNT TO P4 FOR I=4

IF P4>0

? STR(P4,3,0)

?? ' genes with priority = 4 (Secondary analysis:)'

list off fields number,RFEND,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT for I=4

?

ENDIF

COUNT TO P3 FOR I=3

IF P3>0

? STR(P3,3,0)

?? ' genes with priority = 3 (Full insert sequence:)'

list off fields number,RFEND,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT for I=3

?

ENDIF

COUNT TO P2 FOR I=2

IF P2>0

? STR(P2,3,0)

?? ' genes with priority = 2 (Primary analysis complete:)'

list off fields number,RFEND,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT for I=2

?

ENDIF

COUNT TO P1 FOR I=1

IF P1>0

```
? STR(P1,3,0)
?? ' genes with priority = 1 (Primary analysis needed:)'
list off fields number,RFEND,L,D,P,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT for I=1
ENDIF
```

```
*SET PRINT OFF
CLOSE DATABASES
ERASE TEMP1.DBF
ERASE TEMP2.DBF
USE 'SmartGuy\FoxBASE+/Mac:fox files:clones.dbf'
```

```

* COMPRESSION SUBROUTINE FOR ANALYSIS PROGRAMS
USE TEMP1
COUNT TO TOT
REPLACE ALL RFEND WITH 1
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
  IF MARK1 >= TOT
    PACK
    COUNT TO UNIQUE
    SW2=1
  LOOP
ENDIF
GO MARK1
DUP = 1
STORE ENTRY TO TESTA
SW = 0
DO WHILE SW=0 TEST
  SKIP
  STORE ENTRY TO TESTB
  IF TESTA = TESTB
    DELETE
    DUP = DUP+1
  LOOP
ENDIF
GO MARK1
REPLACE RFEND WITH DUP
MARK1 = MARK1+DUP
SW=1
LOOP
ENDDO TEST
LOOP
ENDDO ROLL
*BROWSE
*SET PRINTER ON
SORT ON NUMBER TO TEMP2
USE TEMP2

?? STR(UNIQUE,4,0)
?? ' genes, for a total of '
?? STR(TOT,5,0)
?? ' clones'
? '
      V Coincidence'
list off fields number,RFEND,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I

*SET PRINT OFF
CLOSE DATABASES
ERASE TEMP1.DBF
ERASE TEMP2.DBF
USE "SmartGuy\FoxBASE+Mac:fox files:clones.dbf"

```

```

* COMPRESSION SUBROUTINE FOR ANALYSIS PROGRAMS
USE TEMP1
COUNT TO TOT
REPLACE ALL RFEND WITH 1
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
  IF MARK1 >= TOT
    PACK
    COUNT TO UNIQUE
    COUNT TO NEWGENES FOR D='H'.OR.D='O'
    SW2=1
  LOOP
ENDIF
GO MARK1
DUP = 1
STORE ENTRY TO TESTA
SW = 0
DO WHILE SW=0 TEST
  SKIP
  STORE ENTRY TO TESTB
  IF TESTA = TESTB
    DELETE
    DUP = DUP+1
  LOOP
ENDIF
GO MARK1
REPLACE RFEND WITH DUP
MARK1 = MARK1+DUP
SW=1
LOOP
ENDDO TEST
LOOP
ENDDO ROLL
GO TOP
STORE R TO FUNC
USE "Analysis function.dbf"
LOCATE FOR P=FUNC
REPLACE CLONES WITH TOT
REPLACE GENES WITH UNIQUE
REPLACE NEW WITH NEWGENES.
USE TEMP1
SORT ON RFEND/D TO TEMP2
USE TEMP2
SET HEADING ON
?? STR(UNIQUE,5,0)
?? ' genes, for a total of '
?? STR(TOT,5,0)
?? ' clones'
***
? '          V Coincidence'
list off fields number,RFEND,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I
***
*SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Geneva",12 COLOR 0,0,
*list off fields RFEND,S,DESCRIPTOR

*SET PRINT OFF
CLOSE DATABASES
ERASE TEMP1.DBF
ERASE TEMP2.DBF
USE TEMPDESIG

```



```

* COMPRESSION SUBROUTINE FOR ANALYSIS PROGRAMS
USE TEMP1
COUNT TO TOT
REPLACE ALL RFEND WITH 1
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
  IF MARK1 >= TOT
    PACK
    COUNT TO UNIQUE
    SW2=1
  LOOP
ENDIF
GO MARK1
DUP = 1
STORE ENTRY TO TESTA
SW = 0
DO WHILE SW=0 TEST
  SKIP
  STORE ENTRY TO TESTB
  IF TESTA = TESTB
    DELETE
    DUP = DUP+1
  LOOP
ENDIF
GO MARK1
REPLACE RFEND WITH DUP
MARK1 = MARK1+DUP
SW=1
LOOP
ENDDO TEST
LOOP
ENDDO ROLL
GO TOP
STORE F TO DIST
USE "Analysis distribution.dbf"
LOCATE FOR F=DIST
REPLACE CLONES WITH TOT
REPLACE GENES WITH UNIQUE
USE TEMP1
sort on rfend/d to TEMP2
USE TEMP2
?? STR(UNIQUE,5,0)
?? ' genes, for a total of '
?? STR(TOT,5,0)
?? ' clones'
? '          V Coincidence'
list off fields number,RFEND,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I

*SET PRINT OFF
CLOSE DATABASES
ERASE TEMP1.DEF
ERASE TEMP2.DBF
USE TEMPDESIG

```

```

* COMPRESSION SUBROUTINE FOR ANALYSIS PROGRAMS
USE TEMP1
COUNT TO TOT
REPLACE ALL RFEND WITH 1
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
  IF MARK1 >= TOT
    PACK
    COUNT TO UNIQUE
    SW2=1
  LOOP
  ENDIF
GO MARK1
DUP = 1
STORE ENTRY TO TESTA
SW = 0
DO WHILE SW=0 TEST
  SKIP
  STORE ENTRY TO TESTB
  IF TESTA = TESTB
    DELETE
    DUP = DUP+1
  LOOP
  ENDIF
GO MARK1
REPLACE RFEND WITH DUP
MARK1 = MARK1+DUP
SW=1
LOOP
ENDDO TEST
LOOP
ENDDO ROLL
GO TOP
USE TEMP1
?? STR(UNIQUE,5,0)
?? ' genes, for a total of '
?? STR(TOT,5,0)
?? ' clones'
? '
      V Coincidence'
list off fields number,RFEND,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I

*SET PRINT OFF
CLOSE DATABASES
ERASE TEMP1.DBF
USE TEMPDESIG

```

```

* COMPRESSION SUBROUTINE FOR ANALYSIS PROGRAMS
USE "SmartGuy:FoxBASE+/Mac:fox files:Clones.dbf"
COPY TO TEMP1 FOR
USE TEMP1
COUNT TO IDGENE FOR D='E'.OR.D='O'.OR.D='H'.OR.D='N'.OR.D='R'.OR.D='A'
DELETE FOR D='N'.OR.D='D'.OR.D='A'.OR.D='U'.OR.D='S'.OR.D='M'.OR.D='R'.OR.D='V'
PACK
COUNT TO TOT
REPLACE ALL RFEND WITH 1
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
  IF MARK1 >= TOT
    PACK
    COUNT TO UNIQUE
    SW2=1
    LOOP
  ENDIF
GO MARK1
DUP = 1
STORE ENTRY TO TESTA
SW = 0
DO WHILE SW=0 TEST
  SKIP
  STORE ENTRY TO TESTB
  IF TESTA = TESTB
    DELETE
    DUP = DUP+1
    LOOP
  ENDIF
GO MARK1
REPLACE RFEND WITH DUP
MARK1 = MARK1+DUP
SW=1
LOOP
ENDDO TEST
LOOP
ENDDO ROLL
*BROWSE
*SET PRINTER ON
SORT ON RFEND/D,NUMBER TO TEMP2
USE TEMP2
REPLACE ALL START WITH RFEND/IDGENE*10000
?? STR(UNIQUE,5,0)
?? ' genes, for a total of '
?? STR(TOT,5,0)
?? ' clones'
? ' Coincidence V      V Clones/10000'
set heading off
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Geneva",7 COLOR 0,0,0,
list fields number,RFEND,START,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,INIT,I
*SET PRINT OFF
CLOSE DATABASES
ERASE TEMP1.DBF
ERASE TEMP2.DBF
USE "SmartGuy:FoxBASE+/Mac:fox files:clones.dbf"

```

```

* COMPRESSION SUBROUTINE FOR ANALYSIS PROGRAMS
USE TEMP1
COUNT TO IDGENE FOR D='E'.OR.D='O'.OR.D='H'.OR.D='N'.OR.D='R'.OR.D='A'
DELETE FOR D='N'.OR.D='D'.OR.D='A'.OR.D='U'.OR.D='S'.OR.D='M'.OR.D='R'.OR.D='V'
PACK
COUNT TO TOT
REPLACE ALL RFEND WITH 1
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
  IF MARK1 >= TOT
    PACK
    COUNT TO UNIQUE
    SW2=1
    LOOP
  ENDIF
GO MARK1
DUP = 1
STORE ENTRY TO TESTA
SW = 0
DO WHILE SW=0 TEST
  SKIP
  STORE ENTRY TO TESTB
  IF TESTA = TESTB
    DELETE
    DUP = DUP+1
    LOOP
  ENDIF
GO MARK1
REPLACE RFEND WITH DUP
MARK1 = MARK1+DUP
SW=1
LOOP
ENDDO TEST
LOOP
ENDDO ROLL
*BROWSE
*SET PRINTER ON
SORT ON RFEND/D,NUMBER TO TEMP2
USE TEMP2
REPLACE ALL START WITH RFEND/IDGENE*10000
?? STR(UNIQUE,5,0)
?? ' genes, for a total of '
?? STR(TOT,5,0)
?? ' clones'
? ' Coincidence V          V Clones/10000'
set heading off
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Geneva',7 COLOR 0,0,0,
list fields number,RFEND,START,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,INIT,I
*SET PRINT OFF
CLOSE DATABASES
ERASE TEMP1.DBF
ERASE TEMP2.DBF
USE 'SmartGuy:FoxBASE+/Mac:fox files:clones.dbf'

```

```
USE TEMP1
COUNT TO TOT
?? ' Total of'
?? STR(TOT,4,0)
?? ' clones'
?
*list off fields number,L,D,F,Z,R,C,ENTRY,DESCRIPTOR,LENGTH,RFEND,INIT,I
list off fields number,L,D,F,Z,R,C,ENTRY,DESCRIPTOR
CLOSE DATABASES
ERASE TEMP1.DBF
USE TEMPDESIG
```

```

*Lifescan menu; version 8-7-94
SET TALK OFF
set device to screen
CLEAR
USE "SmartGuy:FoxBASE+/Mac:fox files:clones.dbf"
STORE LUPDATE() TO Update
GO BOTTOM
STORE RECNO() TO cloneno
STORE 6 TO Chooser
DO WHILE .T.
  * Program.: Lifeseq menu.fmt
  * Date..... 1/11/95
  * Version.: FoxBASE+/Mac, revision 1.10
  * Notes..... Format file Lifeseq menu
  *
  SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Geneva",268 COLOR 0,0,
  @ PIXELS 18,126 TO 77,365 STYLE 28479 COLOR 32767,-25600,-1,-16223,-16721,-15725
  @ PIXELS 110,29 TO 188,217 STYLE 3871 COLOR 0,0,-1,-25600,-1,-1
  @ PIXELS 45,161 SAY "LIFESEQ" STYLE 65536 FONT "Geneva",536 COLOR 0,0,-1,-1,7135,5884
  @ PIXELS 36,269 SAY "TM" STYLE 65536 FONT "Geneva",12 COLOR 0,0,-1,-1,7135,5884
  @ PIXELS 63,143 SAY "Molecular Biology Desktop" STYLE 65536 FONT "Helvetica",18 COLOR 0,0,0,
  @ PIXELS 90,252 TO 251,467 STYLE 28447 COLOR 0,0,-1,-25600,-1,-1
  @ PIXELS 117,270 GET Chooser STYLE 65536 FONT "Chicago",12 PICTURE "@*RV Transcript profiles
  @ PIXELS 135,128 SAY Update STYLE 0 FONT "Geneva",12 SIZE 15,79 COLOR 0,0,0,-25600,-1,-1
  @ PIXELS 171,128 SAY cloneno STYLE 0 FONT "Geneva",12 SIZE 15,79 COLOR 0,0,0,-25600,-1,-1
  @ PIXELS 135,44 SAY "Last update:" STYLE 65536 FONT "Geneva",12 COLOR 0,0,-1,-1,-1,-1
  @ PIXELS 171,44 SAY "Total clones:" STYLE 65536 FONT "Geneva",12 COLOR 0,0,-1,-1,-1,-1
  @ PIXELS 45,296 SAY "v1.30" STYLE 65536 FONT "Geneva",782 COLOR 0,0,-1,-1,-1,-1
  *
  * EOF: Lifeseq menu.fmt
  READ
  DO CASE
  CASE Chooser=1
  DO "SmartGuy:FoxBASE+/Mac:fox files:Output programs:Master analysis 3.prg"
  CASE Chooser=2
  DO "SmartGuy:FoxBASE+/Mac:fox files:Output programs:Subtraction 2.prg"
  CASE Chooser=3
  DO "SmartGuy:FoxBASE+/Mac:fox files:Output programs:Northern (single).prg"
  CASE Chooser=4
  USE "Libraries.dbf"
  BROWSE
  CASE Chooser=5
  DO "SmartGuy:FoxBASE+/Mac:fox files:Output programs:See individual clone.prg"
  CASE Chooser=6
  DO "SmartGuy:FoxBASE+/Mac:fox files:Libraries:Output programs:Menu.prg"
  CASE Chooser=7
  CLEAR
  SCREEN 1 OFF
  RETURN
  ENDCASE

  LOOP
ENDDO

```

```

Q1,30 SAY "Database Subset Analysis" STYLE 65536, FONT "Geneva",274 COLOR 0,0,0,-1,-1,-1
?
?
?
?
? date()
?? '
?? TIME()
? 'Clone numbers '
?? STR(INITIATE,6,0)
?? ' through '
?? STR(TERMINATE,6,0)
? 'Libraries: '
IF ENTIRE=1
? 'All libraries'
ENDIF
IF ENTIRE=2
MARK=1
DO WHILE .T.
IF MARK>STOPIT
EXIT
ENDIF
USE SELECTED
GO MARK
? '
?? TRIM(libname)
STORE MARK+1 TO MARK
LOOP
ENDDO
ENDIF
? 'Designations: '
IF Ematch=0 .AND. Hmatch=0 .AND. Omatch=0
?? 'All'
ENDIF
IF Ematch=1
?? 'Exact,'
ENDIF
IF Hmatch=1
?? 'Human,'
ENDIF
IF Omatch=1
?? 'Other sp.'
ENDIF
IF CONDENSE=1
? 'Condensed format analysis'
ENDIF
IF ANAL=1
? 'Sorted by NUMBER'
ENDIF
IF ANAL=2
? 'Sorted by ENTRY'
ENDIF
IF ANAL=3
? 'Arranged by ABUNDANCE'
ENDIF
IF ANAL=4
? 'Sorted by INTEREST'
ENDIF
IF ANAL=5
? 'Arranged by LOCATION'
ENDIF
IF ANAL=6
? 'Arranged by DISTRIBUTION'
ENDIF
IF ANAL=7
? 'Arranged by FUNCTION'

```

```
ENDIF
? 'Total clones represented: '
?? STR(STARTOT,6,0)
? 'Total clones analyzed: '
?? STR(ANALTOT,6,0)
?
?
```



```
USE TEMP1
COUNT TO TOT
?? ' Total of'
?? STR(TOT,4,0)
?? ' clones'
?
*list off fields number,L,D,F,Z,R,C,ENTRY,DESCRIPTOR,LENGTH,RFEND,INIT,I
list off fields number,L,D,F,Z,R,C,ENTRY,DESCRIPTOR
CLOSE DATABASES
ERASE TEMP1.DBF
USE TEMPDESIG
```

```
USE TEMP1
COUNT TO TOT
?? ' Total of'
?? STR(TOT,4,0)
?? ' clones!
?
*list off fields number,L,D,F,Z,R,C,ENTRY,DESCRIPTOR,LENGTH,RFEND,INIT,I
list off fields number,L,D,F,Z,R,C,ENTRY,DESCRIPTOR
CLOSE DATABASES
ERASE TEMP1.DBF
USE TEMPDESIG
```

```

*Northern (single), version 11-25-94
close databases
SET TALK OFF
SET PRINT OFF
SET EXACT OFF
CLEAR
STORE '          ' TO Eobject
STORE '          ' TO Dobject
STORE 0 TO Numb
STORE 0 TO Zog
STORE 1 TO Bail
DO WHILE .T.
* Program.: Northern (single).fmt
* Date.....: 8/ 8/94
* Version.: FoxBASE+/Mac, revision 1.10
* Notes.....: Format file Northern (single)
*
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Geneva",12 COLOR 0,0,0
@ PIXELS 15,81 TO 46,397 STYLE 28447 COLOR 0,0,-1,-25600,-1,-1
@ PIXELS 89,79 TO 192,422 STYLE 28447 COLOR 0,0,0,-25600,-1,-1
@ PIXELS 115,98 SAY "Entry #:" STYLE 65536 FONT "Geneva",12 COLOR 0,0,0,-1,-1,-1
@ PIXELS 115,173 GET Eobject STYLE 0 FONT "Geneva",12 SIZE 15,142 COLOR 0,0,0,-1,-1,-1
@ PIXELS 145,89 SAY "Description" STYLE 65536 FONT "Geneva",12 COLOR 0,0,0,-1,-1,-1
@ PIXELS 145,173 GET Dobject STYLE 0 FONT "Geneva",12 SIZE 15,241 COLOR 0,0,0,-1,-1,-1
@ PIXELS 35,89 SAY "Single Northern search screen" STYLE 65536 FONT "Geneva",274 COLOR 0,0,-
@ PIXELS 220,162 GET Bail STYLE 65536 FONT "Chicago",12 PICTURE "@*R Continue;Bail out" SIZE
@ PIXELS 175,98 SAY "Clone #:" STYLE 65536 FONT "Geneva",12 COLOR 0,0,0,-1,-1,-1
@ PIXELS 175,173 GET Numb STYLE 0 FONT "Geneva",12 SIZE 15,70 COLOR 0,0,0,-1,-1,-1
@ PIXELS 80,152 SAY "Enter any ONE of the following:" STYLE 65536 FONT "Geneva",12 COLOR -1,
*
* EOF: Northern (single).fmt
READ
IF Bail=2
CLEAR
screen 1 off
RETURN
ENDIF
USE "SmartGuy:FoxBASE+/Mac:Fox files:Lookup.dbf"
SET TALK ON

IF Eobject<>'
STORE UPPER(Eobject) to Eobject
SET SAFETY OFF
SORT ON Entry TO "Lookup entry.dbf"
SET SAFETY ON
USE "Lookup entry.dbf"
LOCATE FOR Look=Eobject
IF .NOT.FOUND()
CLEAR
LOOP
ENDIF
BROWSE
STORE Entry TO Searchval
CLOSE DATABASES
ERASE "Lookup entry.dbf"
ENDIF

IF Dobject<>'
SET EXACT OFF
SET SAFETY OFF
SORT ON descriptor TO "Lookup descriptor.dbf"
SET SAFETY On
USE "Lookup descriptor.dbf"
LOCATE FOR UPPER(TRIM(descriptor))=UPPER(TRIM(Dobject))
IF .NOT.FOUND()
CLEAR

```

```

LOOP
ENDIF
BROWSE
STORE Entry TO Searchval
CLOSE DATABASES
ERASE "Lookup descriptor.dbf"
SET EXACT ON
ENDIF

IF Numb<>0
USE "SmartGuy:FoxBASE+/Mac:Fox files:clones.dbf"
GO Numb
BROWSE
STORE Entry TO Searchval
ENDIF

CLEAR
? 'Northern analysis for entry '
?? Searchval
?
? 'Enter Y to proceed'
WAIT TO OK
CLEAR
IF UPPER(OK)<>'Y'
screen 1 off
RETURN
ENDIF

* COMPRESSION SUBROUTINE FOR Library.dbf
? 'Compressing the Libraries file now...'
USE "SmartGuy:FoxBASE+/Mac:Fox files:libraries.dbf"
SET SAFETY OFF
SORT ON library TO "Compressed libraries.dbf"
* FOR entered>0
SET SAFETY ON
USE "Compressed libraries.dbf"
DELETE FOR entered=0
PACK
COUNT TO TOT
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
    IF MARK1 >= TOT
        PACK
        SW2=1
        LOOP
    ENDIF
GO MARK1
STORE library TO TESTA
SKIP
STORE Library TO TESTB
IF TESTA = TESTB
DELETE
ENDIF
MARK1 = MARK1+1
LOOP
ENDDO ROLL

* Northern analysis
CLEAR
? 'Doing the northern now...'
SET TALK ON
USE "SmartGuy:FoxBASE+/Mac:Fox files:clones.dbf"
SET SAFETY OFF
COPY TO "Hits.dbf" FOR entry=searchval
SET SAFETY ON

```

```
CLOSE DATABASES
SELECT 1
USE "Compressed libraries.dbf"
STORE RECCOUNT() TO Entries
SELECT 2
USE "Hits.dbf"
Mark=1
DO WHILE .T.
  SELECT 1
  IF Mark>Entries
    EXIT
  ENDIF
  GO MARK
  STORE library TO Jigger
  SELECT 2
  COUNT TO Zog FOR library=Jigger
  SELECT 1
  REPLACE hits with Zog
  Mark=Mark+1
  LOOP
ENDDO

SELECT 1
BROWSE FIELDS LIBRARY,LIBNAME,ENTERED,HITS AT 0,0
CLEAR
? 'Enter Y to print:'
WAIT TO PRINSET
IF UPPER(PRINSET)='Y'
  SET PRINT ON
  CLEAR
  EJECT
  SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Geneva",14 COLOR 0,0,0
  ? 'DATABASE ENTRIES MATCHING ENTRY '
  ?? Searchval
  ? DATE()
  ?
  SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Geneva",7 COLOR 0,0,0,
  LIST OFF FIELDS library,libname,entered,hits
  ?
  ?
  SELECT 2
  LIST OFF FIELDS NUMBER,LIBRARY,D,S,F,Z,R,ENTRY,DESCRIPTOR,RFSTART,START,RFEND
  SET TALK OFF
  SET PRINT OFF
  ENDIF
CLOSE DATABASES
SET TALK OFF
CLEAR
DO "Test print.prg"
RETURN
```

TABLE 6

library	libname
ADENINB01	Inflamed adenoid
ADRENOR01	Adrenal gland (r)
ADRENOT01	Adrenal gland (T)
AMLBNOT01	AML blast cells (T)
BMAFNOT01	Bone marrow
BMAFNOT02	Bone marrow (T)
CARDNOT01	Cardiac muscle (T)
CHAONOT01	Chin. hamster ovary
CORNNOT01	Corneal stroma
FIBRAGT01	Fibroblast, AT 5
FIBRAGT02	Fibroblast, AT 30
FIBRANT01	Fibroblast, AT
FIBRNGT01	Fibroblast, uv 5
FIBRNGT02	Fibroblast, uv 30
FIBRNOT01	Fibroblast
FIBRNOT02	Fibroblast, normal
HMC1NOT01	Mast cell line HMC-1
HUVELPB01	HUVEC IFN,TNF,LPS
HUVENOB01	HUVEC control
HUVESTB01	HUVEC shear stress
HYPONOB01	Hypothalamus
KIDNNOT01	Kidney (T)
LIVRNOT01	Liver (T)
LUNGNOT01	Lung (T)
MUSCNOT01	Skeletal muscle (T)
OVIDNOB01	Oviduct
PANCNOT01	Pancreas, normal
PITUNOR01	Pituitary (r)
PITUNOT01	Pituitary (T)
PLACNOB01	Placenta
SINTNOT02	Small intestine (T)
SPLNFET01	Spleen+liver, fetal
SPLNNOT02	Spleen (T)
STOMNOT01	Stomach
SYNORAB01	Rheum. synovium
TBLYNOT01	T + B lymphoblast
TESTNOT01	Testis (T)
THP1NOB01	THP-1 control
THP1PEB01	THP phorbol
THP1PLB01	THP-1 phorbol LPS
U937NOT01	U937, monocytic leuk

number	library	d	s	f	z	r	entry	descriptor	rf	atar	atar	rf	end
2304	U937NOT01	E	H	C	C	T	HUMEF1B	Elongation factor 1-beta	0	0	773		
3240	HMC1NOT01	E	H	C	C	T	HUMEF1B	Elongation factor 1-beta	0	370	773		
3259	HMC1NOT01	E	H	C	C	T	HUMEF1B	Elongation factor 1-beta	0	371	773		
4693	HMC1NOT01	E	H	C	C	T	HUMEF1B	Elongation factor 1-beta	0	470	773		
8989	HMC1NOT01	E	H	C	C	T	HUMEF1B	Elongation factor 1-beta	0	327	773		
9139	HMC1NOT01	E	H	C	C	T	HUMEF1B	Elongation factor 1-beta	0	375	773		

WHAT IS CLAIMED IS:

1. A method of analyzing a specimen containing gene transcripts, said method comprising the steps of:
 - (a) producing a library of biological sequences;
 - 5 (b) generating a set of transcript sequences, where each of the transcript sequences in said set is indicative of a different one of the biological sequences of the library;
 - (c) processing the transcript sequences in a
10 programmed computer in which a database of reference transcript sequences indicative of reference biological sequences is stored, to generate an identified sequence value for each of the transcript sequences, where each said identified sequence value is indicative of a sequence
15 annotation and a degree of match between one of the transcript sequences and at least one of the reference transcript sequences; and
 - (d) processing each said identified sequence value to generate final data values indicative of a number of times
20 each identified sequence value is present in the library.
2. The method of claim 1, wherein step (a) includes the steps of:
 - obtaining a mixture of mRNA;
 - making cDNA copies of the mRNA;
 - 25 isolating a representative population of clones transfected with the cDNA and producing therefrom the library of biological sequences.
3. The method of claim 1, wherein the biological sequences are cDNA sequences.
- 30 4. The method of claim 1, wherein the biological sequences are RNA sequences.
5. The method of claim 1, wherein the biological sequences are protein sequences.

6. The method of claim 1, wherein a first value of said degree of match is indicative of an exact match, and a second value of said degree of match is indicative of a non-exact match.

5 7. A method of comparing two specimens containing gene transcripts, said method comprising:

(a) analyzing a first specimen according to the method of claim 1;

10 (b) producing a second library of biological sequences;

(c) generating a second set of transcript sequences, where each of the transcript sequences in said second set is indicative of a different one of the biological sequences of the second library;

15 (d) processing the second set of transcript sequences in said programmed computer to generate a second set of identified sequence values known as further identified sequence values, where each of the further identified sequence values is indicative of a sequence annotation and
20 a degree of match between one of the biological sequences of the second library and at least one of the reference sequences;

(e) processing each said further identified sequence value to generate further final data values indicative of a
25 number of times each further identified sequence value is present in the second library; and

(f) processing the final data values from the first specimen and the further identified sequence values from the second specimen to generate ratios of transcript
30 sequences, each of said ratio values indicative of differences in numbers of gene transcripts between the two specimens.

8. A method of quantifying relative abundance of mRNA in a biological specimen, said method comprising the steps
35 of:

(a) isolating a population of mRNA transcripts from the biological specimen;

(b) identifying genes from which the mRNA was transcribed by a sequence-specific method;

(c) determining numbers of mRNA transcripts corresponding to each of the genes; and

5 (d) using the mRNA transcript numbers to determine the relative abundance of mRNA transcripts within the population of mRNA transcripts.

9. A diagnostic method which comprises producing a gene transcript image, said method comprising the steps of:

10 (a) isolating a population of mRNA transcripts from a biological specimen;

(b) identifying genes from which the mRNA was transcribed by a sequence-specific method;

(c) determining numbers of mRNA transcripts
15 corresponding to each of the genes; and

(d) using the mRNA transcript numbers to determine the relative abundance of mRNA transcripts within the population of mRNA transcripts, where data determining the relative abundance values of mRNA transcripts is the gene
20 transcript image of the biological specimen.

10. The method of claim 9, further comprising:

(e) providing a set of standard normal and diseased gene transcript images; and

(f) comparing the gene transcript image of the
25 biological specimen with the gene transcript images of step (e) to identify at least one of the standard gene transcript images which most closely approximate the gene transcript image of the biological specimen.

11. The method of claim 9, wherein the biological
30 specimen is biopsy tissue, sputum, blood or urine.

12. A method of producing a gene transcript image, said method comprising the steps of

(a) obtaining a mixture of mRNA;

(b) making cDNA copies of the mRNA;

- (c) inserting the cDNA into a suitable vector and using said vector to transfect suitable host strain cells which are plated out and permitted to grow into clones, each clone representing a unique mRNA;
- 5 (d) isolating a representative population of recombinant clones;
- (e) identifying amplified cDNAs from each clone in the population by a sequence-specific method which identifies gene from which the unique mRNA was transcribed;
- 10 (f) determining a number of times each gene is represented within the population of clones as an indication of relative abundance; and
- (g) listing the genes and their relative abundance in order of abundance, thereby producing the gene transcript
- 15 image.

13. The method of claim 12, also including the step of diagnosing disease by:

- repeating steps (a) through (g) on biological specimens from random sample of normal and diseased humans,
- 20 encompassing a variety of diseases, to produce reference sets of normal and diseased gene transcript images;
- obtaining a test specimen from a human, and producing a test gene transcript image by performing steps (a) through (g) on said test specimen;
- 25 comparing the test gene transcript image with the reference sets of gene transcript images; and
- identifying at least one of the reference gene transcript images which most closely approximates the test gene transcript image.

30 14. A computer system for analyzing a library of biological sequences, said system including:

- means for receiving a set of transcript sequences, where each of the transcript sequences is indicative of a different one of the biological sequences of the library;
- 35 and
- means for processing the transcript sequences in the computer system in which a database of reference transcript

sequences indicative of reference biological sequences is stored, wherein the computer is programmed with software for generating an identified sequence value for each of the transcript sequences, where each said identified sequence value is indicative of a sequence annotation and a degree of match between a different one of the biological sequences of the library and at least one of the reference transcript sequences, and for processing each said identified sequence value to generate final data values indicative of a number of times each identified sequence value is present in the library.

15. The system of claim 14, also including:
library generation means for producing the library of biological sequences and generating said set of transcript sequences from said library.

16. The system of claim 15, wherein the library generation means includes:
means for obtaining a mixture of mRNA;
means for making cDNA copies of the mRNA;
means for inserting the cDNA copies into cells and permitting the cells to grow into clones;
means for isolating a representative population of the clones and producing therefrom the library of biological sequences.

SYBASE database Structure

Library Preparation

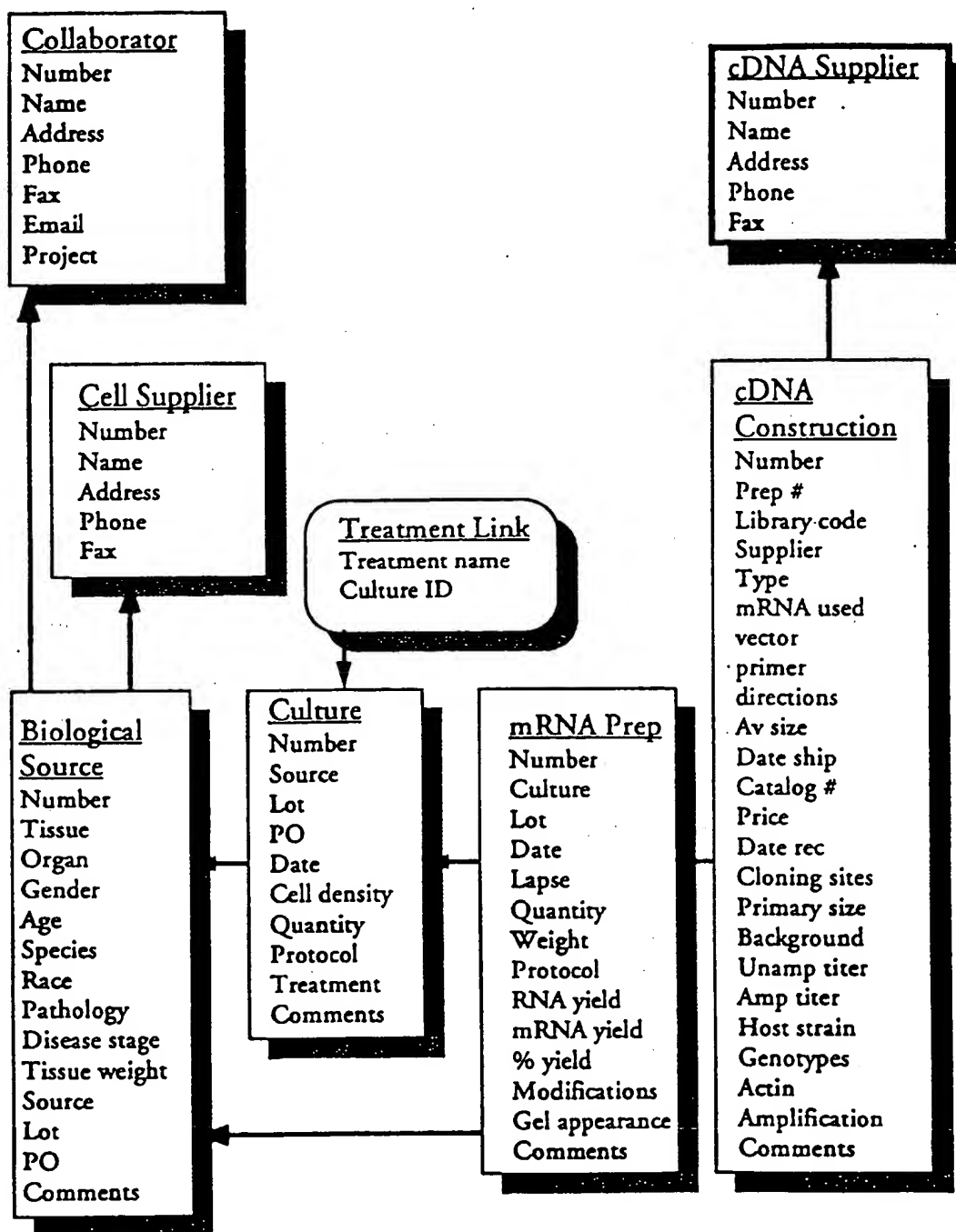


Figure 1

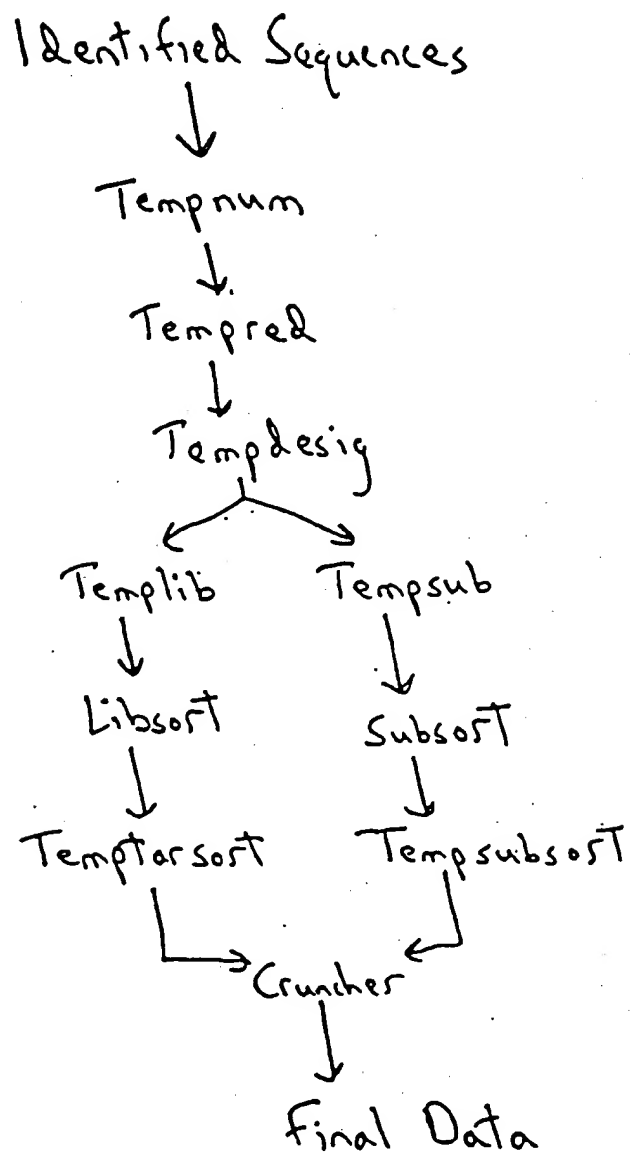


Figure 2

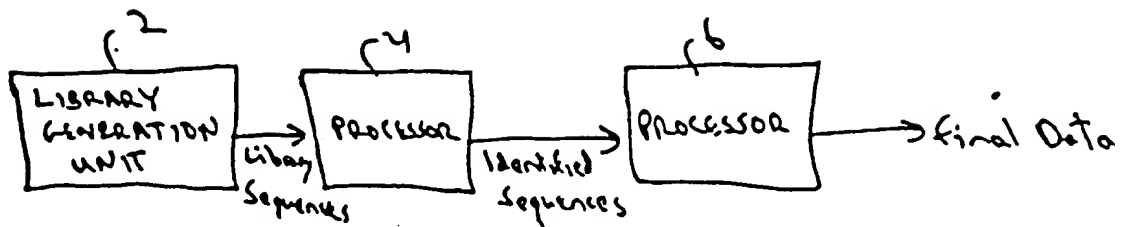


Figure 3

Incyte Bioinformatics Process

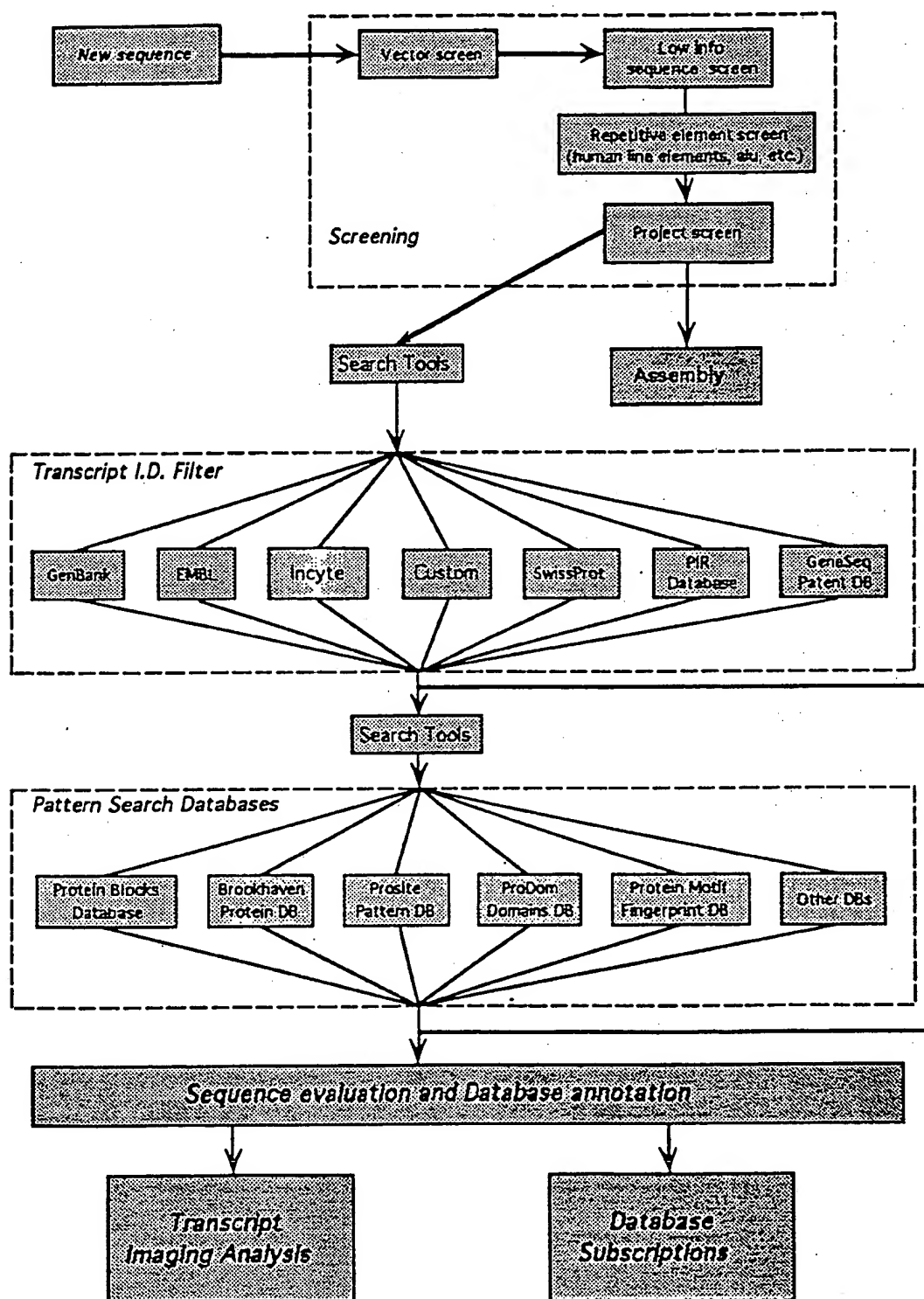


Figure 4

A. CLASSIFICATION OF SUBJECT MATTER

IPC(6) : C12Q 1/68; G06F 15/00

US CL : 435/6; 364/413.02

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 435/6; 364/413.02

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CAS ONLINE, APS, transcript, transcripts, cdan#, mrna#, frequenc?, distribut?, abundanc?

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	IntelliGenetics Suite, Release 5.4, Advanced Training Manual, issued January 1993 by IntelliGenetics, Inc. 700 East El Camino Real, Mountain View, California 94040, United States of America, pages (1-6)-(1-19) and (2-9)-(2-14), see entire document.	15 and 16
Y		----- 1-14
Y	Science, Volume 252, issued 21 June 1991, M.D. Adams et al, "Complementary DNA sequencing: Expressed sequence tags and human genome project", pages 1651-1656, see entire document.	1-16

☒ Further documents are listed in the continuation of Box C.
 ☐ See patent family annex.

Special categories of cited documents:	
A document defining the general state of the art which is not considered to be of particular relevance	*T* later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
E earlier document published on or after the international filing date	*X* document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
L document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	*Y* document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
O document referring to an oral disclosure, use, exhibition or other means	*A* document member of the same patent family
P document published prior to the international filing date but later than the priority date claimed	

Date of the actual completion of the international search

27 APRIL 1995

Date of mailing of the international search report

04 MAY 1995

 Name and mailing address of the ISA/US
 Commissioner of Patents and Trademarks
 Box PCT
 Washington, D.C. 20231

Facsimile No. (703) 305-3230

Authorized officer

JAMES MARTINELL

Telephone No. (703) 308-0196

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US95/01160

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	Nucleic Acids Research, Volume 19, No. 25, issued 1991, E. Hara et al, "Subtractive cDNA cloning using oligo(dT) ₃₀ -latex and PCR: isolation of cDNA clones specific to undifferentiated human embryonal carcinoma cells", pages 7097-7104, see entire document.	1-16
X	Nature Genetics, Volume 2, No. 3, issued November 1992, K. Okubo et al, "Large scale cDNA sequencing for analysis of	1, 3
---	quantitative and qualitative aspects of gene expression", pages	-----
Y	173-179, see narrative text portion of entire document.	2 and 4-16

REPORTS

Ad1p sequence following Ser¹⁰⁸ and occurs within the domain of Ad1p that shows homology with hIDE (14). To delete the complete STE23 sequence and create the *ste23Δ::URA3* mutation, polymerase chain reaction (PCR) primers (5'-TCGGAAGAGCTCAT-TCTTGTCTATTGATATTGCTC- TGATAGATTG-TACTGAGAGTGCAC-3'; and 5'-GCTACAACAGC-GTCACTTGAATGCCCGACATCTTCGACTGT-GGGGTATTTCACACCG-3') were used to amplify the URA3 sequence of pRS316, and the reaction product was transformed into yeast for one-step gene replacement [R. Rothstein, *Methods Enzymol.* 194, 281 (1991)]. To create the *axl1Δ::LEU2* mutation contained on p114, a 5.0-kb *Sal* I fragment from pAXL1 was cloned into pUC19, and an internal 4.0-kb *Hpa* I-Xho I fragment was replaced with a *LEU2* fragment. To construct the *ste23Δ::LEU2* allele (a deletion corresponding to 931 amino acids) carried on p153, a *LEU2* fragment was used to replace the 2.8-kb *Pml* I-Ecl136 II fragment of STE23, which occurs within a 6.2-kb *Hind* III-Bgl II genomic fragment carried on pSP72 (Promega). To create YEpMFA1, a 1.8-kb *Bam* HI fragment containing MFA1, from pKK16 [K. Kuchler, R. E. Stern, J. Thormer, *EMBO J.* 8, 3973 (1989)], was ligated into the *Bam* HI site of YEp351 [J. E. Hill, A. M. Myers, T. J. Koerner, A. Tzagoloff, *Yeast* 2, 163 (1986)].

uct. pC225 is a KS+ (Stratagene) plasmid containing a 0.5-kb *Bam* HI-Sst I fragment from pAXL1. Substitution mutations of the proposed active site of Ad1p were created with the use of pC225 and site-specific mutagenesis involving appropriate synthetic oligonucleotides (*axl1*-H68A, 5'-GTGCTCACAAGGCGT-GCCAAACGGC-3'; *axl1*-E71A, 5'-AAGAATCAT-GTGGCACAAGGCGCGC-3'; and *axl1*-E71D, 5'-AAGAATCATGTGATCACAAGGCGCGC-3'). The mutations were confirmed by sequence analysis. After mutagenesis, the 0.4-kb *Bam* HI-Msc I fragment from the mutagenized pC225 plasmids was transferred into pAXL1 to create a set of pRS316 plasmids carrying different AXL1 alleles, p124 (*axl1*-H68A), p130 (*axl1*-E71A), and p132 (*axl1*-E71D). Similarly, a set of HA-tagged alleles carried on YEp352 were created after replacement of the p151 *Bam* HI-Msc I fragment, to generate p161 (*axl1*-E71A), p162 (*axl1*-

32

N. Davis, T. Favero, C. de Hoog, and S. Kim for comments on the manuscript. Supported by a grant to C.B. from the Natural Sciences and Engineering Research Council of Canada. Support for M.N.A. was from a California Tobacco-Related Disease Research Program postdoctoral fellowship (4FT-0083).

22 June 1995; accepted 21 August 1995

Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray

Mark Schena,* Dari Shalon,*† Ronald W. Davis, Patrick O. Brown‡

A high-capacity system was developed to monitor the expression of many genes in parallel. Microarrays prepared by high-speed robotic printing of complementary DNAs on glass were used for quantitative expression measurements of the corresponding genes. Because of the small format and high density of the arrays, hybridization volumes of 2 microliters could be used that enabled detection of rare transcripts in probe mixtures derived from 2 micrograms of total cellular messenger RNA. Differential expression measurements of 45 *Arabidopsis* genes were made by means of simultaneous, two-color fluorescence hybridization.

The temporal, developmental, topographical, histological, and physiological patterns in which a gene is expressed provide clues to its biological role. The large and expanding database of complementary DNA (cDNA) sequences from many organisms (1) presents the opportunity of defining these patterns at the level of the whole genome.

For these studies, we used the small flowering plant *Arabidopsis thaliana* as a model organism. *Arabidopsis* possesses many advantages for gene expression analysis, including the fact that it has the smallest genome of any higher eukaryote examined to date (2). Forty-five cloned *Arabidopsis* cDNAs (Table 1), including 14 complete sequences and 31 expressed sequence tags (ESTs), were used as gene-specific targets. We obtained the ESTs by selecting cDNA clones at random from an *Arabidopsis* cDNA library. Sequence analysis revealed that 28 of the 31 ESTs matched sequences

in the database (Table 1). Three additional cDNAs from other organisms served as controls in the experiments.

The 48 cDNAs, averaging ~1.0 kb, were amplified with the polymerase chain reaction (PCR) and deposited into individual wells of a 96-well microtiter plate. Each sample was duplicated in two adjacent wells to allow the reproducibility of the arraying and hybridization process to be tested. Samples from the microtiter plate were printed onto glass microscope slides in an area measuring 3.5 mm by 5.5 mm with the use of a high-speed arraying machine (3). The arrays were processed by chemical and heat treatment to attach the DNA sequences to the glass surface and denature them (3). Three arrays, printed in a single lot, were used for the experiments here. A single microtiter plate of PCR products provides sufficient material to print at least 500 arrays.

Fluorescent probes were prepared from total *Arabidopsis* mRNA (4) by a single round of reverse transcription (5). The *Arabidopsis* mRNA was supplemented with human acetylcholine receptor (AChR) mRNA at a dilution of 1:10,000 (w/w) before cDNA synthesis, to provide an internal standard for calibration (5). The resulting fluorescently labeled cDNA mixture was hybridized to an array at high stringency (6) and scanned

M. Schena and R. W. Davis, Department of Biochemistry, Beckman Center, Stanford University Medical Center, Stanford, CA 94305, USA.

D. Shalon and P. O. Brown, Department of Biochemistry and Howard Hughes Medical Institute, Beckman Center, Stanford University Medical Center, Stanford, CA 94305, USA.

*These authors contributed equally to this work.

†Present address: Syntel, Palo Alto, CA 94303, USA.

‡To whom correspondence should be addressed. E-mail: pbrown@cmsgn.stanford.edu

24. J. Chant and I. Herskowitz, *Cell* 65, 1203 (1991).
25. B. W. Matthews, *Acc. Chem. Res.* 21, 333 (1988).
26. K. Kuchler, H. G. Dohlman, J. Thormer, *J. Cell Biol.* 120, 1203 (1993); R. Koling and C. P. Hollenberg, *EMBO J.* 13, 3281 (1994); C. Berkower, D. Loeyza, S. Michaels, *Mol. Biol. Cell* 5, 1185 (1994).
27. A. Bender and J. R. Pringle, *Proc. Natl. Acad. Sci. U.S.A.* 86, 9976 (1989); J. Chant, K. Corrado, J. R. Pringle, I. Herskowitz, *Cell* 65, 1213 (1991); S. Powers, E. Gonzales, T. Christensen, J. Cubert, D. Broek, *ibid.*, p. 1225; H. O. Park, J. Chant, I. Herskowitz, *Nature* 365, 269 (1993); J. Chant, *Trends Genet.* 10, 328 (1994); and J. R. Pringle, *J. Cell Biol.* 129, 751 (1995); J. Chant, M. Mischke, E. Mitchell, I. Herskowitz, J. R. Pringle, *ibid.*, p. 767.
28. G. F. Sprague Jr., *Methods. Enzymol.* 194, 77 (1991).
29. Single-letter abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.
30. A W303 1A derivative, SY2625 (*MATa ura3-1 leu2-3, 112 trp1-1 ade2-1 can1-100 ss11Δ mfa2Δ::FUS1-lacZ his3Δ::FUS1-HIS3*), was the parent strain for the mutant search. SY2625 derivatives for the mating assays, secreted pheromone assays, and the pulse-chase experiments included the following strains: Y49 (*ste22-1*), Y115 (*mfa1Δ::LEU2*), Y142 (*axl1::URA3*), Y173 (*axl1Δ::LEU2*), Y220 (*axl1::URA3 ste23Δ::URA3*), Y231 (*ste23Δ::URA3*), Y231 (*axl1Δ::LEU2 ste23Δ::LEU2*), and Y233 (*ste23Δ::LEU2*). *MATa* derivatives of SY2625 included the following strains: Y189 (*SY2625 made MATa*), Y278 (*ste22-1*), Y185 (*mfa1Δ::LEU2*), Y186 (*axl1Δ::LEU2*), and Y187 (*axl1::URA3*). The EG123 (*MATa leu2 ura3 trp1 can1 his4*) genetic background was used to create a set of strains for analysis of bud site selection. EG123 derivatives included the following strains: Y175 (*axl1Δ::LEU2*), Y223 (*axl1::URA3*), Y234 (*ste23Δ::LEU2*), and Y272 (*axl1Δ::LEU2 ste23Δ::LEU2*). *MATa* derivatives of EG123 included the following strains: Y214 (EG123 made *MATa*) and Y293 (*axl1Δ::LEU2*). All strains were generated by means of standard genetic or molecular methods involving the appropriate constructs (23). In particular, the *axl1 ste23* double mutant strains were created by crossing of the appropriate *MATa ste23* and *MATa axl1* mutants, followed by sporulation of the resultant diploid and isolation of the double mutant from nonparental d-type tetrads. Gene disruptions were confirmed with either PCR or Southern (DNA) analysis.
31. p129 is a YEp352 [J. E. Hill, A. M. Myers, T. J. Koerner, A. Tzagoloff, *Yeast* 2, 163 (1986)] plasmid containing a 5.5-kb *Sal* I fragment of pAXL1. p151 was derived from p129 by insertion of a linker at the *Bgl* II site within AXL1, which led to an in-frame insertion of the hemagglutinin (HA) epitope (DYPPDYPPYA) (29) between amino acids 854 and 855 of the AXL1 prod-

with a laser (3). A high-sensitivity scan gave signals that saturated the detector at nearly all of the *Arabidopsis* target sites (Fig. 1A). Calibration relative to the AChR mRNA standard (Fig. 1A) established a sensitivity limit of $\sim 1:50,000$. No detectable hybridization was observed to either the rat glucocorticoid receptor (Fig. 1A) or the yeast TRP4 (Fig. 1A) targets even at the highest scanning sensitivity. A moderate-sensitivity scan

of the same array allowed linear detection of the more abundant transcripts (Fig. 1B). Quantitation of both scans revealed a range of expression levels spanning three orders of magnitude for the 45 genes tested (Table 2). RNA blots (7) for several genes (Fig. 2) corroborated the expression levels measured with the microarray to within a factor of 5 (Table 2).

Differential gene expression was investi-

gated with a simultaneous, two-color hybridization scheme, which served to minimize experimental variation inherent in the comparison of independent hybridizations. Fluorescent probes were prepared from two mRNA sources with the use of reverse transcriptase in the presence of fluorescein- and lissamine-labeled nucleotide analogs, respectively (5). The two probes were then mixed together in equal proportions, hybridized to a single array, and scanned separately for fluorescein and lissamine emission after independent excitation of the two fluorophores (3).

To test whether overexpression of a single gene could be detected in a pool of total *Arabidopsis* mRNA, we used a microarray to analyze a transgenic line overexpressing the single transcription factor HAT4 (8). Fluorescent probes representing mRNA from wild-type and HAT4-transgenic plants were labeled with fluorescein and lissamine, respectively; the two probes were then mixed and hybridized to a single array. An intense hybridization signal was observed at the position of the HAT4 cDNA in the lissamine-specific scan (Fig. 1D), but not in the fluorescein-specific scan of the same array (Fig. 1C). Calibration with AChR mRNA added to the fluorescein and lissamine cDNA synthesis reactions at dilutions of 1:10,000 (Fig. 1C) and 1:100 (Fig. 1D), respectively, revealed a 50-fold elevation of HAT4 mRNA in the transgenic line relative to its abundance in wild-type plants (Table 2). This magnitude of HAT4 overexpression matched that inferred from the Northern (RNA) analysis within a factor of 2 (Fig. 2 and Table 2). Expression of all the other genes monitored on the array differed by less than a factor of 5 between HAT4-transgenic and wild-type plants (Fig. 1, C

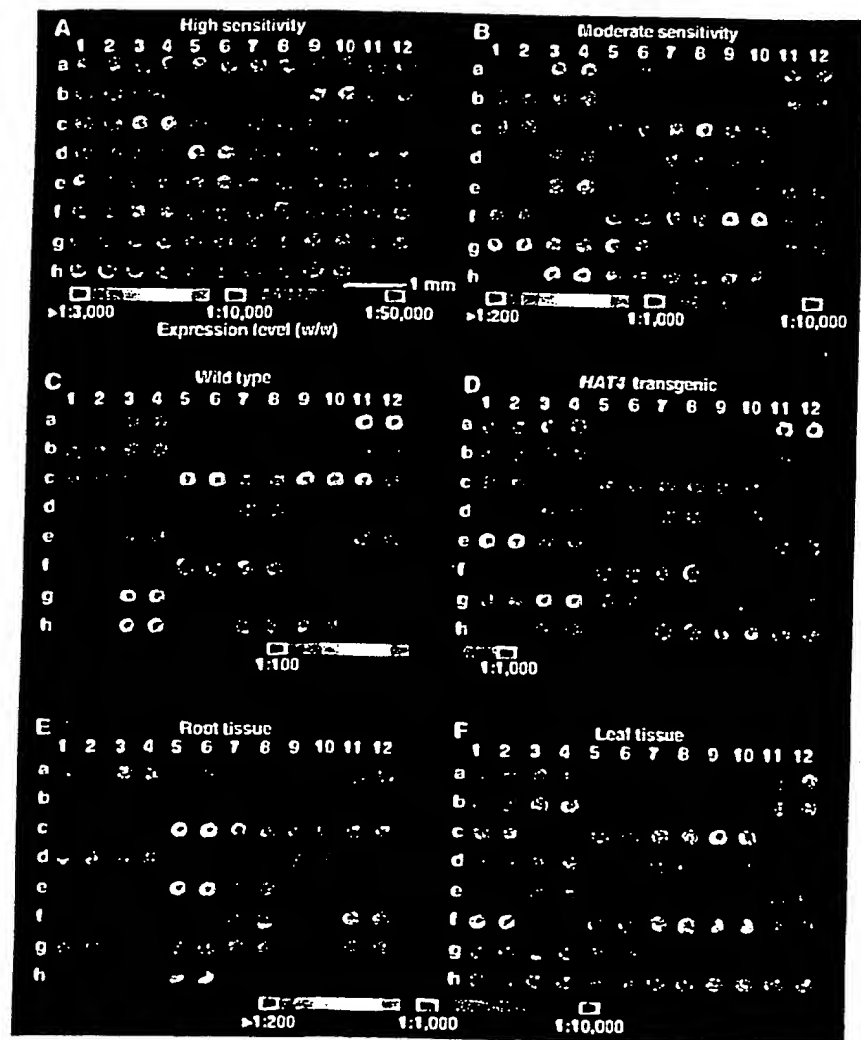


Fig. 1. Gene expression monitored with the use of cDNA microarrays. Fluorescent scans represented in pseudocolor correspond to hybridization intensities. Color bars were calibrated from the signal obtained with the use of known concentrations of human AChR mRNA in independent experiments. Numbers and letters on the axes mark the position of each cDNA. (A) High-sensitivity fluorescein scan after hybridization with fluorescein-labeled cDNA derived from wild-type plants. (B) Same array as in (A) but scanned at moderate sensitivity. (C and D) A single array was probed with a 1:1 mixture of fluorescein-labeled cDNA from wild-type plants and lissamine-labeled cDNA from HAT4-transgenic plants. The single array was then scanned successively to detect the fluorescein fluorescence corresponding to mRNA from wild-type plants (C) and the lissamine fluorescence corresponding to mRNA from HAT4-transgenic plants (D). (E and F) A single array was probed with a 1:1 mixture of fluorescein-labeled cDNA from root tissue and lissamine-labeled cDNA from leaf tissue. The single array was then scanned successively to detect the fluorescein fluorescence corresponding to mRNAs expressed in roots (E) and the lissamine fluorescence corresponding to mRNAs expressed in leaves (F).

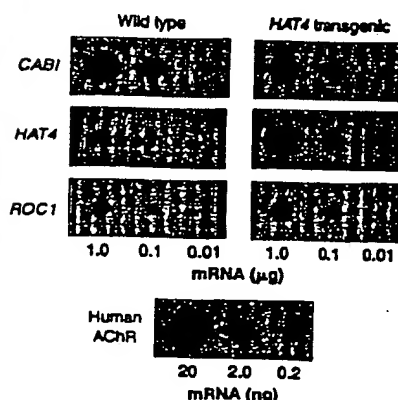


Fig. 2. Gene expression monitored with RNA (Northern) blot analysis. Designated amounts of mRNA from wild-type and HAT4-transgenic plants were spotted onto nylon membranes and probed with the cDNAs indicated. Purified human AChR mRNA was used for calibration.

and D, and Table 2). Hybridization of fluorescein-labeled glucocorticoid receptor cDNA (Fig. 1C) and lissamine-labeled TRP4 cDNA (Fig. 1D) verified the presence of the negative control targets and the lack of optical cross talk between the two fluorophores.

To explore a more complex alteration in expression patterns, we performed a second two-color hybridization experiment with fluorescein- and lissamine-labeled probes prepared from root and leaf mRNA, respectively. The scanning sensitivities for the two fluorophores were normalized by matching the signals resulting from AChR

mRNA, which was added to both cDNA synthesis reactions at a dilution of 1:1000 (Fig. 1, E and F). A comparison of the scans revealed widespread differences in gene expression between root and leaf tissue (Fig. 1, E and F). The mRNA from the light-regulated CAB1 gene was ~500-fold more abundant in leaf (Fig. 1F) than in root tissue (Fig. 1E). The expression of 26 other genes differed between root and leaf tissue by more than a factor of 5 (Fig. 1, E and F).

The HAT4-transgenic line we examined has elongated hypocotyls, early flowering, poor germination, and altered pigmentation (8). Although changes in expression were

observed for HAT4, large changes in expression were not observed for any of the other 44 genes we examined. This was somewhat surprising, particularly because comparative analysis of leaf and root tissue identified 27 differentially expressed genes. Analysis of an expanded set of genes may be required to identify genes whose expression changes upon HAT4 overexpression; alternatively, a comparison of mRNA populations from specific tissues of wild-type and HAT4-transgenic plants may allow identification of downstream genes.

At the current density of robotic printing, it is feasible to scale up the fabrication process to produce arrays containing 20,000 cDNA targets. At this density, a single array would be sufficient to provide gene-specific targets encompassing nearly the entire repertoire of expressed genes in the *Arabidopsis* genome (2). The availability of 20,274 ESTs from *Arabidopsis* (1, 9) would provide a rich source of templates for such studies.

The estimated 100,000 genes in the human genome (10) exceeds the number of *Arabidopsis* genes by a factor of 5 (2). This modest increase in complexity suggests that similar cDNA microarrays, prepared from the rapidly growing repertoire of human ESTs (1), could be used to determine the expression patterns of tens of thousands of human genes in diverse cell types. Coupling an amplification strategy to the reverse transcription reaction (11) could make it feasible to monitor expression even in minute tissue samples. A wide variety of acute and chronic physiological and pathological conditions might lead to characteristic changes in the patterns of gene expression in peripheral blood cells or other easily sampled tissues. In concert with cDNA microarrays for monitoring complex expression patterns, these tissues might therefore serve as sensitive *in vivo* sensors for clinical diagnosis. Microarrays of cDNAs could thus provide a useful link between human gene sequences and clinical medicine.

Table 2. Gene expression monitoring by microarray and RNA blot analyses; tg, HAT4-transgenic. See Table 1 for additional gene information. Expression levels (w/w) were calibrated with the use of known amounts of human AChR mRNA. Values for the microarray were determined from microarray scans (Fig. 1); values for the RNA blot were determined from RNA blots (Fig. 2).

Gene	Expression level (w/w)	
	Microarray	RNA blot
CAB1	1:48	1:83
CAB1 (tg)	1:120	1:150
HAT4	1:8300	1:6300
HAT4 (tg)	1:150	1:210
ROC1	1:1200	1:1800
ROC1 (tg)	1:260	1:1300

Table 1. Sequences contained on the cDNA microarray. Shown is the position, the known or putative function, and the accession number of each cDNA in the microarray (Fig. 1). All but three of the ESTs used in this study matched a sequence in the database. NADH, reduced form of nicotinamide adenine dinucleotide; ATPase, adenosine triphosphatase; GTP, guanosine triphosphate.

Position	cDNA	Function	Accession number
a1, 2	AChR	Human AChR	-
a3, 4	EST3	Actin	H36236
a5, 6	EST6	NADH dehydrogenase	Z27010
a7, 8	AAC1	Actin 1	M20016
a8, 10	EST12	Unknown	U36594†
a11, 12	EST13	Actin	T45783
b1, 2	CAB1	Chlorophyll a/b binding	M85150
b3, 4	EST17	Phosphoglycerate kinase	T44490
b5, 6	G44	Gibberellic acid biosynthesis	L37126
b7, 8	EST19	Unknown	U36595†
b9, 10	GBF-1	G-box binding factor 1	X63894
b11, 12	EST23	Elongation factor	X52256
c1, 2	EST29	Aldolase	T04477
c3, 4	GBF-2	G-box binding factor 2	X63895
c5, 6	EST34	Chloroplast protease	R87034
c7, 8	EST35	Unknown	T14152
c9, 10	EST41	Catalase	T22720
c11, 12	rGR	Rat glucocorticoid receptor	M14053
d1, 2	EST42	Unknown	U36596†
d3, 4	EST45	ATPase	J04185
d5, 6	HAT1	Homeobox-leucine zipper 1	U09332
d7, 8	EST46	Light harvesting complex	T04063
d9, 10	EST49	Unknown	T76267
d11, 12	HAT2	Homeobox-leucine zipper 2	U09335
e1, 2	HAT4	Homeobox-leucine zipper 4	M90394
e3, 4	EST50	Phosphoribulokinase	T04344
e5, 6	HAT5	Homeobox-leucine zipper 5	M90416
e7, 8	EST51	Unknown	Z33675
e9, 10	HAT22	Homeobox-leucine zipper 22	U09336
e11, 12	EST52	Oxygen evolving	T21749
f1, 2	EST59	Unknown	Z34607
f3, 4	KNAT1	Knotted-like homeobox 1	U14174
f5, 6	EST60	RuBisCO small subunit	X14564
f7, 8	EST69	Translation elongation factor	T42799
f9, 10	PPH1	Protein phosphatase 1	U34803
f11, 12	EST70	Unknown	T44621
g1, 2	EST75	Chloroplast protease	T43698
g3, 4	EST78	Unknown	R65481
g5, 6	ROC1	Cyclophilin	L14844
g7, 8	EST82	GTP binding	X59152
g9, 10	EST83	Unknown	Z33795
g11, 12	EST84	Unknown	T45278
h1, 2	EST91	Unknown	T13832
h3, 4	EST96	Unknown	R64816
h5, 6	SAR1	Synaptobrevin	M90418
h7, 8	EST100	Light harvesting complex	Z18205
h9, 10	EST103	Light harvesting complex	X03909
h11, 12	TRP4	Yeast tryptophan biosynthesis	X04273

*Proprietary sequence of Stratagene (La Jolla, California).

†No match in the database; novel EST.

REFERENCES AND NOTES

1. The current EST database (dbEST release 091495) from the National Center for Biotechnology Information (Bethesda, MD) contains a total of 322,225 entries, including 255,645 from the human genome and 21,044 from Arabidopsis. Access is available via the World Wide Web (<http://www.ncbi.nlm.nih.gov>).
2. E. M. Meyerowitz and R. E. Pruitt, *Science* 228, 1214 (1985); R. E. Pruitt and E. M. Meyerowitz, *J. Mol. Biol.* 187, 169 (1986); L. Hwang et al., *Plant J.* 1, 367 (1991); P. Jarvis et al., *Plant Mol. Biol.* 24, 685 (1994); L. Le Guan et al., *Mol. Gen. Genet.* 245, 390 (1994).
3. D. Shalon, thesis, Stanford University (1995); and P. O. Brown, in preparation. Microarrays were fabricated on poly-L-lysine-coated microscope slides (Sigma) with a custom-built arraying machine fitted with one printing tip. The tip loaded 1 μ l of PCR product (0.5 mg/ml) from 96-well microtiter plates and deposited \sim 0.005 μ l per slide on 40 slides at a spacing of 500 μ m. The printed slides were rehydrated for 2 hours in a humid chamber, snap-dried at 100°C for 1 min, rinsed in 0.1% SDS, and treated with 0.05% succinic anhydride prepared in buffer consisting of 50% 1-methyl-2-pyrrolidinone and 50% boric acid. The cDNA on the slides was denatured in distilled water for 2 min at 90°C immediately before use. Microarrays were scanned with a laser fluorescent scanner that contained a computer-controlled XY stage and a microscope objective. A mixed gas, multiline laser allowed sequential excitation of the two fluorophores. Emitted light was split according to wavelength and detected with two photomultiplier tubes. Signals were read into a PC with the use of a 12-bit analog-to-digital board. Additional details of microarray fabrication and use may be obtained by means of e-mail (pbrown@cmgm.stanford.edu).
4. F. M. Ausubel et al., Eds., *Current Protocols in Molecular Biology* (Greene & Wiley Interscience, New York, 1994), pp. 4.3.1–4.3.4.
5. Polyadenylated [poly(A)⁺] mRNA was prepared from total RNA with the use of Oligotex-dT resin (Qiagen). Reverse transcription (RT) reactions were carried out with a Stratascript RT-PCR kit (Stratagene) modified as follows: 50- μ l reactions contained 0.1 μ g/ μ l of Arabidopsis mRNA, 0.1 ng/ μ l of human AChR mRNA, 0.05 μ g/ μ l of oligo(dT) (21-mer), 1 \times first strand buffer, 0.03 U/ μ l of ribonuclease block, 500 μ M deoxyadenosine triphosphate (dATP), 500 μ M deoxyguanosine triphosphate, 500 μ M dTTP, 40 μ M deoxycytosine triphosphate (dCTP), 40 μ M fluorescein-12-dCTP (or Issamine-5-dCTP), and 0.03 U/ μ l of Stratascript reverse transcriptase. Reactions were incubated for 60 min at 37°C, precipitated with ethanol, and resuspended in 10 μ l of TE (10 mM Tris-HCl and 1 mM EDTA, pH 8.0). Samples were then heated for 3 min at 94°C and chilled on ice. The RNA was degraded by adding 0.25 μ l of 10 N NaOH followed by a 10-min incubation at 37°C. The samples were neutralized by addition of 2.5 μ l of 1 M Tris-Cl (pH 8.0) and 0.25 μ l of 10 N HCl and precipitated with ethanol. Pellets were washed with 70% ethanol, dried to completion in a speedvac, resuspended in 10 μ l of H₂O, and reduced to 3.0 μ l in a speedvac. Fluorescent nucleotide analogs were obtained from New England Nuclear (DuPont).
6. Hybridization reactions contained 1.0 μ l of fluorescent cDNA synthesis product (5) and 1.0 μ l of hybridization buffer (10 \times saline sodium citrate (SSC) and 0.2% SDS). The 2.0- μ l probe mixtures were aliquoted onto the microarray surface and covered with cover slips (12 mm round). Arrays were transferred to a hybridization chamber (3) and incubated for 18 hours at 65°C. Arrays were washed for 5 min at room temperature (25°C) in low-stringency wash buffer (1 \times SSC and 0.1% SDS), then for 10 min at room temperature in high-stringency wash buffer (0.1 \times SSC and 0.1% SDS). Arrays were scanned in 0.1 \times SSC with the use of a fluorescence laser-scanning device (3).
7. Samples of poly(A)⁺ mRNA (4, 5) were spotted onto nylon membranes (Nytan) and crosslinked with ultraviolet light with the use of a Stratalink 1800 (Stratagene). Probes were prepared by random priming with the use of a Prime-It II kit (Stratagene) in the presence of [³²P]dATP. Hybridizations were carried out according to the instructions of the manufacturer. Quantitation was performed on a Phosphorimager (Molecular Dynamics).
8. M. Schena and R. W. Davis, *Proc. Natl. Acad. Sci. U.S.A.* 89, 3894 (1992); M. Schena, A. M. Lloyd, R. W. Davis, *Genes Dev.* 7, 367 (1993); M. Schena and R. W. Davis, *Proc. Natl. Acad. Sci. U.S.A.* 91, 8393 (1994).
9. H. Hofte et al., *Plant J.* 4, 1051 (1993); T. Newman et al., *Plant Physiol.* 106, 1241 (1994).
10. N. E. Morton, *Proc. Natl. Acad. Sci. U.S.A.* 88, 7474 (1991); E. D. Green and R. H. Waterston, *J. Am. Med. Assoc.* 266, 1966 (1991); C. Bellanne-Chantelot, *Cell* 70, 1059 (1992); D. R. Cox et al., *Science* 265, 2031 (1994).
11. E. S. Kawasaki et al., *Proc. Natl. Acad. Sci. U.S.A.* 85, 5698 (1988).
12. The laser fluorescent scanner was designed and fabricated in collaboration with S. Smith of Stanford University. Scanner and analysis software was developed by R. X. Xia. The succinic anhydride reaction was suggested by J. Mulligan and J. Van Ness of Darwin Molecular Corporation. Thanks to S. Theologis, C. Somerville, K. Yamamoto, and members of the laboratories of R.W.D. and P.O.B. for critical comments. Supported by the Howard Hughes Medical Institute and by grants from NIH (R21HG00450) (P.O.B.) and R37AG00198 (R.W.D.) and from NSF (MCB9106011) (R.W.D.) and by an NSF graduate fellowship (D.S.). P.O.B. is an assistant investigator of the Howard Hughes Medical Institute.

11 August 1995; accepted 22 September 1995

Gene Therapy in Peripheral Blood Lymphocytes and Bone Marrow for ADA⁻ Immunodeficient Patients

Claudio Bordignon,* Luigi D. Notarangelo, Nadia Nobili, Giuliana Ferrari, Giulia Casorati, Paola Panina, Evelina Mazzolari, Daniela Maggioni, Claudia Rossi, Paolo Servida, Alberto G. Ugazio, Fulvio Mavilio

Adenosine deaminase (ADA) deficiency results in severe combined immunodeficiency, the first genetic disorder treated by gene therapy. Two different retroviral vectors were used to transfer ex vivo the human ADA minigene into bone marrow cells and peripheral blood lymphocytes from two patients undergoing exogenous enzyme replacement therapy. After 2 years of treatment, long-term survival of T and B lymphocytes, marrow cells, and granulocytes expressing the transferred ADA gene was demonstrated and resulted in normalization of the immune repertoire and restoration of cellular and humoral immunity. After discontinuation of treatment, T lymphocytes, derived from transduced peripheral blood lymphocytes, were progressively replaced by marrow-derived T cells in both patients. These results indicate successful gene transfer into long-lasting progenitor cells, producing a functional multilineage progeny.

Severe combined immunodeficiency associated with inherited deficiency of ADA (1) is usually fatal unless affected children are kept in protective isolation or the immune system is reconstituted by bone marrow transplantation from a human leukocyte antigen (HLA)-identical sibling donor (2). This is the therapy of choice, although it is available only for a minority of patients. In recent years, other forms of therapy have been developed, including transplants from haploidentical donors (3, 4), exogenous enzyme replacement (5), and somatic-cell gene therapy (6–9).

We previously reported a preclinical model in which ADA gene transfer and expression

successfully restored immune functions in human ADA-deficient (ADA⁻) peripheral blood lymphocytes (PBLs) in immunodeficient mice in vivo (10, 11). On the basis of these preclinical results, the clinical application of gene therapy for the treatment of ADA⁻ SCID (severe combined immunodeficiency disease) patients who previously failed exogenous enzyme replacement therapy was approved by our Institutional Ethical Committees and by the Italian National Committee for Bioethics (12). In addition to evaluating the safety and efficacy of the gene therapy procedure, the aim of the study was to define the relative role of PBLs and hematopoietic stem cells in the long-term reconstitution of immune functions after retroviral vector-mediated ADA gene transfer. For this purpose, two structurally identical vectors expressing the human ADA complementary DNA (cDNA), distinguishable by the presence of alternative restriction sites in a nonfunctional region of the viral long-terminal repeat (LTR), were used to transduce PBLs and bone marrow (BM) cells independently. This procedure allowed identification of the origin of

C. Bordignon, N. Nobili, G. Ferrari, D. Maggioni, C. Rossi, P. Servida, F. Mavilio, Telethon Gene Therapy Program for Genetic Diseases, DIBIT, Istituto Scientifico H. S. Raffaele, Milan, Italy.
L. D. Notarangelo, E. Mazzolari, A. G. Ugazio, Department of Pediatrics, University of Brescia Medical School, Brescia, Italy.
G. Casorati, Unità di Immunochimica, DIBIT, Istituto Scientifico H. S. Raffaele, Milan, Italy.
P. Panina, Roche Milano Ricerche, Milan, Italy.

*To whom correspondence should be addressed.



INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification ⁶ : G01N 33/543, 33/68	A1	(11) International Publication Number: WO 95/35505 (43) International Publication Date: 28 December 1995 (28.12.95)
(21) International Application Number: PCT/US95/07659 (22) International Filing Date: 16 June 1995 (16.06.95) (30) Priority Data: 08/261,388 17 June 1994 (17.06.94) US 08/477,809 7 June 1995 (07.06.95) US (71) Applicant: THE BOARD OF TRUSTEES OF THE LELAND STANFORD JUNIOR UNIVERSITY [US/US]; Stanford, CA 94305 (US). (72) Inventors: SHALON, Tidhar, Dari; 364 Fletcher Drive, Atherton, CA 94027 (US). BROWN, Patričk, O.; 76 Peter Courtis Circle, Stanford, CA 94305 (US). (74) Agent: DEHLINGER, Peter, J.; Dehlinger & Associates, P.O. Box 60850, Palo Alto, CA 94306-1546 (US).		(81) Designated States: AU, CA, JP, European patent (AT, BE, CH, DE, DK, ES, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE). Published <i>With international search report.</i>
(54) Title: METHOD AND APPARATUS FOR FABRICATING MICROARRAYS OF BIOLOGICAL SAMPLES (57) Abstract A method and apparatus for forming microarrays of biological samples on a support are disclosed. The method involves dispensing a known volume of a reagent at each of a selected array position, by tapping a capillary dispenser on the support under conditions effective to draw a defined volume of liquid onto the support. The apparatus is designed to produce a microarray of such regions in an automated fashion.		

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AT	Austria	GB	United Kingdom	MR	Mauritania
AU	Australia	GE	Georgia	MW	Malawi
BB	Barbados	GN	Guinea	NE	Niger
BE	Belgium	GR	Greece	NL	Netherlands
BF	Burkina Faso	HU	Hungary	NO	Norway
BG	Bulgaria	IE	Ireland	NZ	New Zealand
BJ	Benin	IT	Italy	PL	Poland
BR	Brazil	JP	Japan	PT	Portugal
BY	Belarus	KE	Kenya	RO	Romania
CA	Canada	KG	Kyrgyzstan	RU	Russian Federation
CF	Central African Republic	KP	Democratic People's Republic of Korea	SD	Sudan
CG	Congo	KR	Republic of Korea	SE	Sweden
CH	Switzerland	KZ	Kazakhstan	SI	Slovenia
CI	Côte d'Ivoire	LI	Liechtenstein	SK	Slovakia
CM	Cameroon	LK	Sri Lanka	SN	Senegal
CN	China	LU	Luxembourg	TD	Chad
CS	Czechoslovakia	LV	Latvia	TG	Togo
CZ	Czech Republic	MC	Monaco	TJ	Tajikistan
DE	Germany	MD	Republic of Moldova	TT	Trinidad and Tobago
DK	Denmark	MG	Madagascar	UA	Ukraine
ES	Spain	ML	Mali	US	United States of America
FI	Finland	MN	Mongolia	UZ	Uzbekistan
FR	France			VN	Viet Nam
GA	Gabon				

**METHOD AND APPARATUS FOR FABRICATING
MICROARRAYS OF BIOLOGICAL SAMPLES**

Field of the Invention

5 This invention relates to a method and apparatus
for fabricating microarrays of biological samples for
large scale screening assays, such as arrays of DNA
samples to be used in DNA hybridization assays for
genetic research and diagnostic applications.

10

References

- Abouzied, et al., *Journal of AOAC International*
77(2):495-500 (1994).
- Bohlander, et al., *Genomics* 13:1322-1324 (1992).
- 15 Drmanac, et al., *Science* 260:1649-1652 (1993).
- Fodor, et al., *Science* 251:767-773 (1991).
- Khrapko, et al., *DNA Sequence* 1:375-388 (1991).
- Kuriyama, et al., AN ISFET BIOSENSOR, APPLIED BIOSENSORS
(Donald Wise, Ed.), Butterworths, pp. 93-114 (1989).
- 20 Lehrach, et al., HYBRIDIZATION FINGERPRINTING IN GENOME
MAPPING AND SEQUENCING, GENOME ANALYSIS, VOL 1 (Davies and
Tilgham, Eds.), Cold Spring Harbor Press, pp. 39-81
(1990).
- Maniatis, et al., MOLECULAR CLONING, A LABORATORY
25 MANUAL, Cold Spring Harbor Press (1989).
- Nelson, et al., *Nature Genetics* 4:11-18 (1993).

Pirrung, et al., U.S. Patent No. 5,143,854 (1992).

Riles, et al., *Genetics* 134:81-150 (1993).

Schena, M. et al., *Proc. Nat. Acad. Sci. USA*
89:3894-3898 (1992).

5 Southern, et al., *Genomics* 13:1008-1017 (1992).

Background of the Invention

A variety of methods are currently available for making arrays of biological macromolecules, such as
10 arrays of nucleic acid molecules or proteins. One method for making ordered arrays of DNA on a porous membrane is a "dot blot" approach. In this method, a vacuum manifold transfers a plurality, e.g., 96, aqueous samples of DNA from 3 millimeter diameter wells
15 to a porous membrane. A common variant of this procedure is a "slot-blot" method in which the wells have highly-elongated oval shapes.

The DNA is immobilized on the porous membrane by baking the membrane or exposing it to UV radiation.
20 This is a manual procedure practical for making one array at a time and usually limited to 96 samples per array. "Dot-blot" procedures are therefore inadequate for applications in which many thousand samples must be determined.

25 A more efficient technique employed for making ordered arrays of genomic fragments uses an array of pins dipped into the wells, e.g., the 96 wells of a microtitre plate, for transferring an array of samples to a substrate, such as a porous membrane. One array
30 includes pins that are designed to spot a membrane in a staggered fashion, for creating an array of 9216 spots in a 22 x 22 cm area (Lehrach, et al., 1990). A limitation with this approach is that the volume of DNA spotted in each pixel of each array is highly variable.

In addition, the number of arrays that can be made with each dipping is usually quite small.

An alternate method of creating ordered arrays of nucleic acid sequences is described by Pirrung, et al. (1992), and also by Fodor, et al. (1991). The method involves synthesizing different nucleic acid sequences at different discrete regions of a support. This method employs elaborate synthetic schemes, and is generally limited to relatively short nucleic acid sample, e.g., less than 20 bases. A related method has been described by Southern, et al. (1992).

Khrapko, et al. (1991) describes a method of making an oligonucleotide matrix by spotting DNA onto a thin layer of polyacrylamide. The spotting is done manually with a micropipette.

None of the methods or devices described in the prior art are designed for mass fabrication of microarrays characterized by (i) a large number of micro-sized assay regions separated by a distance of 50-200 microns or less, and (ii) a well-defined amount, typically in the picomole range, of analyte associated with each region of the array.

Furthermore, current technology is directed at performing such assays one at a time to a single array of DNA molecules. For example, the most common method for performing DNA hybridizations to arrays spotted onto porous membrane involves sealing the membrane in a plastic bag (Maniatis, et al., 1989) or a rotating glass cylinder (Robbins Scientific) with the labeled hybridization probe inside the sealed chamber. For arrays made on non-porous surfaces, such as a microscope slide, each array is incubated with the labeled hybridization probe sealed under a coverslip. These techniques require a separate sealed chamber for

each array which makes the screening and handling of many such arrays inconvenient and time intensive.

Abouzied, et al. (1994) describes a method of printing horizontal lines of antibodies on a nitrocellulose membrane and separating regions of the membrane with vertical stripes of a hydrophobic material. Each vertical stripe is then reacted with a different antigen and the reaction between the immobilized antibody and an antigen is detected using a standard ELISA colorimetric technique. Abouzied's technique makes it possible to screen many one-dimensional arrays simultaneously on a single sheet of nitrocellulose. Abouzied makes the nitrocellulose somewhat hydrophobic using a line drawn with PAP Pen (Research Products International). However Abouzied does not describe a technology that is capable of completely sealing the pores of the nitrocellulose. The pores of the nitrocellulose are still physically open and so the assay reagents can leak through the hydrophobic barrier during extended high temperature incubations or in the presence of detergents which makes the Abouzied technique unacceptable for DNA hybridization assays.

Porous membranes with printed patterns of hydrophilic/hydrophobic regions exist for applications such as ordered arrays of bacteria colonies. QA Life Sciences (San Diego CA) makes such a membrane with a grid pattern printed on it. However, this membrane has the same disadvantage as the Abouzied technique since reagents can still flow between the gridded arrays making them unusable for separate DNA hybridization assays.

Pall Corporation make a 96-well plate with a porous filter heat sealed to the bottom of the plate. These plates are capable of containing different

reagents in each well without cross-contamination. However, each well is intended to hold only one target element whereas the invention described here makes a microarray of many biomolecules in each subdivided region of the solid support. Furthermore, the 96 well plates are at least 1 cm thick and prevent the use of the device for many colorimetric, fluorescent and radioactive detection formats which require that the membrane lie flat against the detection surface. The invention described here requires no further processing after the assay step since the barriers elements are shallow and do not interfere with the detection step thereby greatly increasing convenience.

Hyseq Corporation has described a method of making an "array of arrays" on a non-porous solid support for use with their sequencing by hybridization technique. The method described by Hyseq involves modifying the chemistry of the solid support material to form a hydrophobic grid pattern where each subdivided region contains a microarray of biomolecules. Hyseq's flat hydrophobic pattern does not make use of physical blocking as an additional means of preventing cross contamination.

25 Summary of the Invention

The invention includes, in one aspect, a method of forming a microarray of analyte-assay regions on a solid support, where each region in the array has a known amount of a selected, analyte-specific reagent. The method involves first loading a solution of a selected analyte-specific reagent in a reagent-dispensing device having an elongate capillary channel (i) formed by spaced-apart, coextensive elongate members, (ii) adapted to hold a quantity of the reagent solution and (iii) having a tip region at which aqueous

solution in the channel forms a meniscus. The channel is preferably formed by a pair of spaced-apart tapered elements.

5 The tip of the dispensing device is tapped against a solid support at a defined position on the support surface with an impulse effective to break the meniscus in the capillary channel deposit a selected volume of solution on the surface, preferably a selected volume in the range 0.01 to 100 nl. The two steps are
10 repeated until the desired array is formed.

The method may be practiced in forming a plurality of such arrays, where the solution-depositing step is are applied to a selected position on each of a plurality of solid supports at each repeat cycle.

15 The dispensing device may be loaded with a new solution, by the steps of (i) dipping the capillary channel of the device in a wash solution, (ii) removing wash solution drawn into the capillary channel, and (iii) dipping the capillary channel into the new
20 reagent solution.

Also included in the invention is an automated apparatus for forming a microarray of analyte-assay regions on a plurality of solid supports, where each region in the array has a known amount of a selected,
25 analyte-specific reagent. The apparatus has a holder for holding, at known positions, a plurality of planar supports, and a reagent dispensing device of the type described above.

The apparatus further includes positioning
30 structure for positioning the dispensing device at a selected array position with respect to a support in said holder, and dispensing structure for moving the dispensing device into tapping engagement against a support with a selected impulse effective to deposit a

selected volume on the support, e.g., a selected volume in the volume range 0.01 to 100 nl.

The positioning and dispensing structures are controlled by a control unit in the apparatus. The unit operates to (i) place the dispensing device at a loading station, (ii) move the capillary channel in the device into a selected reagent at the loading station, to load the dispensing device with the reagent, and (iii) dispense the reagent at a defined array position on each of the supports on said holder. The unit may further operate, at the end of a dispensing cycle, to wash the dispensing device by (i) placing the dispensing device at a washing station, (ii) moving the capillary channel in the device into a wash fluid, to load the dispensing device with the fluid, and (iii) remove the wash fluid prior to loading the dispensing device with a fresh selected reagent.

The dispensing device in the apparatus may be one of a plurality of such devices which are carried on the arm for dispensing different analyte assay reagents at selected spaced array positions.

In another aspect, the invention includes a substrate with a surface having a microarray of at least 10^3 distinct polynucleotide or polypeptide biopolymers in a surface area of less than about 1 cm^2 . Each distinct biopolymer (i) is disposed at a separate, defined position in said array, (ii) has a length of at least 50 subunits, and (iii) is present in a defined amount between about 0.1 femtomoles and 100 nanomoles.

In one embodiment, the surface is glass slide surface coated with a polycationic polymer, such as polylysine, and the biopolymers are polynucleotides. In another embodiment, the substrate has a water-impermeable backing, a water-permeable film formed on

the backing, and a grid formed on the film. The grid is composed of intersecting water-impervious grid elements extending from said backing to positions raised above the surface of said film, and partitions the film into a plurality of water-impervious cells. A biopolymer array is formed within each well.

More generally, there is provided a substrate for use in detecting binding of labeled polynucleotides to one or more of a plurality different-sequence, immobilized polynucleotides. The substrate includes, in one aspect, a glass support, a coating of a polycationic polymer, such as polylysine, on said surface of the support, and an array of distinct polynucleotides electrostatically bound non-covalently to said coating, where each distinct biopolymer is disposed at a separate, defined position in a surface array of polynucleotides.

In another aspect, the substrate includes a water-impermeable backing, a water-permeable film formed on the backing, and a grid formed on the film, where the grid is composed of intersecting water-impervious grid elements extending from the backing to positions raised above the surface of the film, forming a plurality of cells. A biopolymer array is formed within each cell.

Also forming part of the invention is a method of detecting differential expression of each of a plurality of genes in a first cell type, with respect to expression of the same genes in a second cell type. In practicing the method, there is first produced fluorescent-labeled cDNA's from mRNA's isolated from the two cells types, where the cDNA'S from the first and second cells are labeled with first and second different fluorescent reporters.

A mixture of the labeled cDNA's from the two cell types is added to an array of polynucleotides

representing a plurality of known genes derived from the two cell types, under conditions that result in hybridization of the cDNA's to complementary-sequence polynucleotides in the array. The array is then
5 examined by fluorescence under fluorescence excitation conditions in which (i) polynucleotides in the array that are hybridized predominantly to cDNA's derived from one of the first and second cell types give a distinct first or second fluorescence emission color,
10 respectively, and (ii) polynucleotides in the array that are hybridized to substantially equal numbers of cDNA's derived from the first and second cell types give a distinct combined fluorescence emission color, respectively. The relative expression of known genes
15 in the two cell types can then be determined by the observed fluorescence emission color of each spot.

These and other objects and features of the invention will become more fully apparent when the following detailed description of the invention is read
20 in conjunction with the accompanying figures.

Brief Description of the Drawings

Fig. 1 is a side view of a reagent-dispensing device having a open-capillary dispensing head
25 constructed for use in one embodiment of the invention;

Figs. 2A-2C illustrate steps in the delivery of a fixed-volume bead on a hydrophobic surface employing the dispensing head from Fig. 1, in accordance with one embodiment of the method of the invention;

30 Fig. 3 shows a portion of a two-dimensional array of analyte-assay regions constructed according to the method of the invention;

Fig. 4 is a planar view showing components of an automated apparatus for forming arrays in accordance
35 with the invention.

Fig. 5 shows a fluorescent image of an actual 20 × 20 array of 400 fluorescently-labeled DNA samples immobilized on a poly-l-lysine coated slide, where the total area covered by the 400 element array is 16 square millimeters;

Fig. 6 is a fluorescent image of a 1.8 cm × 1.8 cm microarray containing lambda clones with yeast inserts, the fluorescent signal arising from the hybridization to the array with approximately half the yeast genome labeled with a green fluorophore and the other half with a red fluorophore;

Fig. 7 shows the translation of the hybridization image of Fig. 6 into a karyotype of the yeast genome, where the elements of Fig.-6 microarray contain yeast DNA sequences that have been previously physically mapped in the yeast genome;

Fig. 8 show a fluorescent image of a 0.5 cm × 0.5 cm microarray of 24 cDNA clones, where the microarray was hybridized simultaneously with total cDNA from wild type *Arabidopsis* plant labeled with a green fluorophore and total cDNA from a transgenic *Arabidopsis* plant labeled with a red fluorophore, and the arrow points to the cDNA clone representing the gene introduced into the transgenic *Arabidopsis* plant;

Fig. 9 shows a plan view of substrate having an array of cells formed by barrier elements in the form of a grid;

Fig. 10 shows an enlarged plan view of one of the cells in the substrate in Fig. 9, showing an array of polynucleotide regions in the cell;

Fig. 11 is an enlarged sectional view of the substrate in Fig. 9, taken along a section line in that figure; and

Fig. 12 is a scanned image of a 3 cm × 3 cm nitrocellulose solid support containing four identical

arrays of M13 clones in each of four quadrants, where each quadrant was hybridized simultaneously to a different oligonucleotide using an open face hybridization method.

5

Detailed Description of the Invention

I. Definitions

Unless indicated otherwise, the terms defined below have the following meanings:

10 "Ligand" refers to one member of a ligand/anti-ligand binding pair. The ligand may be, for example, one of the nucleic acid strands in a complementary, hybridized nucleic acid duplex binding pair; an effector molecule in an effector/receptor binding pair;
15 or an antigen in an antigen/antibody or antigen/antibody fragment binding pair.

"Antiligand" refers to the opposite member of a ligand/anti-ligand binding pair. The antiligand may be the other of the nucleic acid strands in a
20 complementary, hybridized nucleic acid duplex binding pair; the receptor molecule in an effector/receptor binding pair; or an antibody or antibody fragment molecule in antigen/antibody or antigen/antibody fragment binding pair, respectively.

25 "Analyte" or "analyte molecule" refers to a molecule, typically a macromolecule, such as a polynucleotide or polypeptide, whose presence, amount, and/or identity are to be determined. The analyte is one member of a ligand/anti-ligand pair.

30 "Analyte-specific assay reagent" refers to a molecule effective to bind specifically to an analyte molecule. The reagent is the opposite member of a ligand/anti-ligand binding pair.

An "array of regions on a solid support" is a
35 linear or two-dimensional array of preferably discrete

regions, each having a finite area, formed on the surface of a solid support.

A "microarray" is an array of regions having a density of discrete regions of at least about $100/\text{cm}^2$, and preferably at least about $1000/\text{cm}^2$. The regions in a microarray have typical dimensions, e.g., diameters, in the range of between about 10-250 μm , and are separated from other regions in the array by about the same distance.

A support surface is "hydrophobic" if a aqueous-medium droplet applied to the surface does not spread out substantially beyond the area size of the applied droplet. That is, the surface acts to prevent spreading of the droplet applied to the surface by hydrophobic interaction with the droplet.

A "meniscus" means a concave or convex surface that forms on the bottom of a liquid in a channel as a result of the surface tension of the liquid.

"Distinct biopolymers", as applied to the biopolymers forming a microarray, means an array member which is distinct from other array members on the basis of a different biopolymer sequence, and/or different concentrations of the same or distinct biopolymers, and/or different mixtures of distinct or different-concentration biopolymers. Thus an array of "distinct polynucleotides" means an array containing, as its members, (i) distinct polynucleotides, which may have a defined amount in each member, (ii) different, graded concentrations of given-sequence polynucleotides, and/or (iii) different-composition mixtures of two or more distinct polynucleotides.

"Cell type" means a cell from a given source, e.g., a tissue, or organ, or a cell in a given state of

differentiation, or a cell associated with a given pathology or genetic makeup.

II. Method of Microarray Formation

5 This section describes a method of forming a microarray of analyte-assay regions on a solid support or substrate, where each region in the array has a known amount of a selected, analyte-specific reagent.

10 Fig. 1 illustrates, in a partially schematic view, a reagent-dispensing device 10 useful in practicing the method. The device generally includes a reagent dispenser 12 having an elongate open capillary channel 14 adapted to hold a quantity of the reagent solution, such as indicated at 16, as will be described below.

15 The capillary channel is formed by a pair of spaced-apart, coextensive, elongate members 12a, 12b which are tapered toward one another and converge at a tip or tip region 18 at the lower end of the channel. More generally, the open channel is formed by at least two

20 elongate, spaced-apart members adapted to hold a quantity of reagent solutions and having a tip region at which aqueous solution in the channel forms a meniscus, such as the concave meniscus illustrated at 20 in Fig. 2A. The advantages of the open channel

25 construction of the dispenser are discussed below.

 With continued reference to Fig. 1, the dispenser device also includes structure for moving the dispenser rapidly toward and away from a support surface, for effecting deposition of a known amount of solution in

30 the dispenser on a support, as will be described below with reference to Figs. 2A-2C. In the embodiment shown, this structure includes a solenoid 22 which is activatable to draw a solenoid piston 24 rapidly downwardly, then release the piston, e.g., under spring

35 bias, to a normal, raised position, as shown. The

dispenser is carried on the piston by a connecting member 26, as shown. The just-described moving structure is also referred to herein as dispensing means for moving the dispenser into engagement with a solid support, for dispensing a known volume of fluid on the support.

The dispensing device just described is carried on an arm 28 that may be moved either linearly or in an x-y plane to position the dispenser at a selected deposition position, as will be described.

Figs. 2A-2C illustrate the method of depositing a known amount of reagent solution in the just-described dispenser on the surface of a solid support, such as the support indicated at 30. The support is a polymer, glass, or other solid-material support having a surface indicated at 31.

In one general embodiment, the surface is a relatively hydrophilic, i.e., wettable surface, such as a surface having native, bound or covalently attached charged groups. On such surface described below is a glass surface having an absorbed layer of a polycationic polymer, such as poly-l-lysine.

In another embodiment, the surface has or is formed to have a relatively hydrophobic character, i.e., one that causes aqueous medium deposited on the surface to bead. A variety of known hydrophobic polymers, such as polystyrene, polypropylene, or polyethylene have desired hydrophobic properties, as do glass and a variety of lubricant or other hydrophobic films that may be applied to the support surface.

Initially, the dispenser is loaded with a selected analyte-specific reagent solution, such as by dipping the dispenser tip, after washing, into a solution of the reagent, and allowing filling by capillary flow into the dispenser channel. The dispenser is now moved

to a selected position with respect to a support surface, placing the dispenser tip directly above the support-surface position at which the reagent is to be deposited. This movement takes place with the
5 dispenser tip in its raised position, as seen in Fig. 2A, where the tip is typically at least several 1-5 mm above the surface of the substrate.

With the dispenser so positioned, solenoid 22 is now activated to cause the dispenser tip to move
10 rapidly toward and away from the substrate surface, making momentary contact with the surface, in effect, tapping the tip of the dispenser against the support surface. The tapping movement of the tip against the surface acts to break the liquid meniscus in the tip
15 channel, bringing the liquid in the tip into contact with the support surface. This, in turn, produces a flowing of the liquid into the capillary space between the tip and the surface, acting to draw liquid out of the dispenser channel, as seen in Fig. 2B.

20 Fig. 2C shows flow of fluid from the tip onto the support surface, which in this case is a hydrophobic surface. The figure illustrates that liquid continues to flow from the dispenser onto the support surface until it forms a liquid bead 32. At a given bead size,
25 i.e., volume, the tendency of liquid to flow onto the surface will be balanced by the hydrophobic surface interaction of the bead with the support surface, which acts to limit the total bead area on the surface, and by the surface tension of the droplet, which tends
30 toward a given bead curvature. At this point, a given bead volume will have formed, and continued contact of the dispenser tip with the bead, as the dispenser tip is being withdrawn, will have little or no effect on bead volume.

For liquid-dispensing on a more hydrophilic surface, the liquid will have less of a tendency to bead, and the dispensed volume will be more sensitive to the total dwell time of the dispenser tip in the immediate vicinity of the support surface, e.g., the positions illustrated in Figs. 2B and 2C.

The desired deposition volume, i.e., bead volume, formed by this method is preferably in the range 2 pl (picoliters) to 2 nl (nanoliters), although volumes as high as 100 nl or more may be dispensed. It will be appreciated that the selected dispensed volume will depend on (i) the "footprint" of the dispenser tip, i.e., the size of the area spanned by the tip, (ii) the hydrophobicity of the support surface, and (iii) the time of contact with and rate of withdrawal of the tip from the support surface. In addition, bead size may be reduced by increasing the viscosity of the medium, effectively reducing the flow time of liquid from the dispenser onto the support surface. The drop size may be further constrained by depositing the drop in a hydrophilic region surrounded by a hydrophobic grid pattern on the support surface.

In a typical embodiment, the dispenser tip is tapped rapidly against the support surface, with a total residence time in contact with the support of less than about 1 msec, and a rate of upward travel from the surface of about 10 cm/sec.

Assuming that the bead that forms on contact with the surface is a hemispherical bead, with a diameter approximately equal to the width of the dispenser tip, as shown in Fig. 2C, the volume of the bead formed in relation to dispenser tip width (d) is given in Table 1 below. As seen, the volume of the bead ranges between 2 pl to 2 nl as the width size is increased from about 20 to 200 μm .

Table 1

d	Volume (nl)
20 μm	2×10^{-3}
50 μm	3.1×10^{-2}
100 μm	2.5×10^{-1}
200 μm	2

5
10 At a given tip size, bead volume can be reduced in a controlled fashion by increasing surface hydrophobicity, reducing time of contact of the tip with the surface, increasing rate of movement of the tip away from the surface, and/or increasing the
15 viscosity of the medium. Once these parameters are fixed, a selected deposition volume in the desired pl to nl range can be achieved in a repeatable fashion.

After depositing a bead at one selected location on a support, the tip is typically moved to a
20 corresponding position on a second support, a droplet is deposited at that position, and this process is repeated until a liquid droplet of the reagent has been deposited at a selected position on each of a plurality of supports.

25 The tip is then washed to remove the reagent liquid, filled with another reagent liquid and this reagent is now deposited at each another array position on each of the supports. In one embodiment, the tip is washed and refilled by the steps of (i) dipping the
30 capillary channel of the device in a wash solution, (ii) removing wash solution drawn into the capillary channel, and (iii) dipping the capillary channel into the new reagent solution.

From the foregoing, it will be appreciated that
35 the tweezers-like, open-capillary dispenser tip

provides the advantages that (i) the open channel of the tip facilitates rapid, efficient washing and drying before reloading the tip with a new reagent, (ii) passive capillary action can load the sample directly from a standard microwell plate while retaining sufficient sample in the open capillary reservoir for the printing of numerous arrays, (iii) open capillaries are less prone to clogging than closed capillaries, and (iv) open capillaries do not require a perfectly faced bottom surface for fluid delivery.

A portion of a microarray 36 formed on the surface 38 of a solid support 40 in accordance with the method just described is shown in Fig. 3. The array is formed of a plurality of analyte-specific reagent regions, such as regions 42, where each region may include a different analyte-specific reagent. As indicated above, the diameter of each region is preferably between about 20-200 μm . The spacing between each region and its closest (non-diagonal) neighbor, measured from center-to-center (indicated at 44), is preferably in the range of about 20-400 μm . Thus, for example, an array having a center-to-center spacing of about 250 μm contains about 40 regions/cm or 1,600 regions/cm². After formation of the array, the support is treated to evaporate the liquid of the droplet forming each region, to leave a desired array of dried, relatively flat regions. This drying may be done by heating or under vacuum.

In some cases, it is desired to first rehydrate the droplets containing the analyte reagents to allow for more time for adsorption to the solid support. It is also possible to spot out the analyte reagents in a humid environment so that droplets do not dry until the arraying operation is complete.

III. Automated Apparatus for Forming Arrays

In another aspect, the invention includes an automated apparatus for forming an array of analyte-assay regions on a solid support, where each region in the array has a known amount of a selected, analyte-specific reagent.

The apparatus is shown in planar, and partially schematic view in Fig. 4. A dispenser device 72 in the apparatus has the basic construction described above with respect to Fig. 1, and includes a dispenser 74 having an open-capillary channel terminating at a tip, substantially as shown in Figs. 1 and 2A-2C.

The dispenser is mounted in the device for movement toward and away from a dispensing position at which the tip of the dispenser taps a support surface, to dispense a selected volume of reagent solution, as described above. This movement is effected by a solenoid 76 as described above. Solenoid 76 is under the control of a control unit 77 whose operation will be described below. The solenoid is also referred to herein as dispensing means for moving the device into tapping engagement with a support, when the device is positioned at a defined array position with respect to that support.

The dispenser device is carried on an arm 74 which is threadedly mounted on a worm screw 80 driven (rotated) in a desired direction by a stepper motor 82 also under the control of unit 77. At its left end in the figure screw 80 is carried in a sleeve 84 for rotation about the screw axis. At its other end, the screw is mounted to the drive shaft of the stepper motor, which in turn is carried on a sleeve 86. The dispenser device, worm screw, the two sleeves mounting the worm screw, and the stepper motor used in moving the device in the "x" (horizontal) direction in the

figure form what is referred to here collectively as a displacement assembly 86.

The displacement assembly is constructed to produce precise, micro-range movement in the direction of the screw, i.e., along an x axis in the figure. In one mode, the assembly functions to move the dispenser in x-axis increments having a selected distance in the range 5-25 μm . In another mode, the dispenser unit may be moved in precise x-axis increments of several microns or more, for positioning the dispenser at associated positions on adjacent supports, as will be described below.

The displacement assembly, in turn, is mounted for movement in the "y" (vertical) axis of the figure, for positioning the dispenser at a selected y axis position. The structure mounting the assembly includes a fixed rod 88 mounted rigidly between a pair of frame bars 90, 92, and a worm screw 94 mounted for rotation between a pair of frame bars 96, 98. The worm screw is driven (rotated) by a stepper motor 100 which operates under the control of unit 77. The motor is mounted on bar 96, as shown.

The structure just described, including worm screw 94 and motor 100, is constructed to produce precise, micro-range movement in the direction of the screw, i.e., along an y axis in the figure. As above, the structure functions in one mode to move the dispenser in y-axis increments having a selected distance in the range 5-250 μm , and in a second mode, to move the dispenser in precise y-axis increments of several microns (μm) or more, for positioning the dispenser at associated positions on adjacent supports.

The displacement assembly and structure for moving this assembly in the y axis are referred to herein collectively as positioning means for positioning the

dispensing device at a selected array position with respect to a support.

A holder 102 in the apparatus functions to hold a plurality of supports, such as supports 104 on which
5 the microarrays of reagent regions are to be formed by the apparatus. The holder provides a number of recessed slots, such as slot 106, which receive the supports, and position them at precise selected positions with respect to the frame bars on which the
10 dispenser moving means is mounted.

As noted above, the control unit in the device functions to actuate the two stepper motors and dispenser solenoid in a sequence designed for automated operation of the apparatus in forming a selected
15 microarray of reagent regions on each of a plurality of supports.

The control unit is constructed, according to conventional microprocessor control principles, to provide appropriate signals to each of the solenoid and
20 each of the stepper motors, in a given timed sequence and for appropriate signalling time. The construction of the unit, and the settings that are selected by the user to achieve a desired array pattern, will be understood from the following description of a typical
25 apparatus operation.

Initially, one or more supports are placed in one or more slots in the holder. The dispenser is then moved to a position directly above a well (not shown) containing a solution of the first reagent to be
30 dispensed on the support(s). The dispenser solenoid is actuated now to lower the dispenser tip into this well, causing the capillary channel in the dispenser to fill. Motors 82, 100 are now actuated to position the dispenser at a selected array position at the first of
35 the supports. Solenoid actuation of the dispenser is

then effective to dispense a selected-volume droplet of that reagent at this location. As noted above, this operation is effective to dispense a selected volume preferably between 2 pl and 2 nl of the reagent solution.

The dispenser is now moved to the corresponding position at an adjacent support and a similar volume of the solution is dispensed at this position. The process is repeated until the reagent has been dispensed at this preselected corresponding position on each of the supports.

Where it is desired to dispense a single reagent at more than two array positions on a support, the dispenser may be moved to different array positions at each support, before moving the dispenser to a new support, or solution can be dispensed at individual positions on each support, at one selected position, then the cycle repeated for each new array position.

To dispense the next reagent, the dispenser is positioned over a wash solution (not shown), and the dispenser tip is dipped in and out of this solution until the reagent solution has been substantially washed from the tip. Solution can be removed from the tip, after each dipping, by vacuum, compressed air spray, sponge, or the like.

The dispenser tip is now dipped in a second reagent well, and the filled tip is moved to a second selected array position in the first support. The process of dispensing reagent at each of the corresponding second-array positions is then carried as above. This process is repeated until an entire microarray of reagent solutions on each of the supports has been formed.

IV. Microarray Substrate

This section describes embodiments of a substrate having a microarray of biological polymers carried on the substrate surface. Subsection A describes a multi-cell substrate, each cell of which contains a microarray, and preferably an identical microarray, of distinct biopolymers, such as distinct polynucleotides, formed on a porous surface. Subsection B describes a microarray of distinct polynucleotides bound on a glass slide coated with a polycationic polymer.

A. Multi-Cell Substrate

Fig. 9 illustrates, in plan view, a substrate 110 constructed according to the invention. The substrate has an 8 × 12 rectangular array 112 of cells, such as cells 114, 116, formed on the substrate surface. With reference to Fig. 10, each cell, such as cell 114, in turn supports a microarray 118 of distinct biopolymers, such as polypeptides or polynucleotides at known, addressable regions of the microarray. Two such regions forming the microarray are indicated at 120, and correspond to regions, such as regions 42, forming the microarray of distinct biopolymers shown in Fig. 3.

The 96-cell array shown in Fig. 9 has typically array dimensions between about 12 and 244 mm in width and 8 and 400 mm in length, with the cells in the array having width and length dimension of 1/12 and 1/8 the array width and length dimensions, respectively, i.e., between about 1 and 20 in width and 1 and 50 mm in length.

The construction of substrate is shown cross-sectionally in Fig. 11, which is an enlarged sectional view taken along view line 124 in Fig. 9. The substrate includes a water-impermeable backing 126, such as a glass slide or rigid polymer sheet. Formed on the surface of the backing is a water-permeable film

128. The film is formed of a porous membrane material, such as nitrocellulose membrane, or a porous web material, such as a nylon, polypropylene, or PVDF porous polymer material. The thickness of the film is preferably between about 10 and 1000 μm . The film may be applied to the backing by spraying or coating uncured material on the backing, or by applying a preformed membrane to the backing. The backing and film may be obtained as a preformed unit from commercial source, e.g., a plastic-backed nitrocellulose film available from Schleicher and Schuell Corporation.

With continued reference to Fig. 11, the film-covered surface in the substrate is partitioned into a desired array of cells by water-impermeable grid lines, such as lines 130, 132, which have infiltrated the film down to the level of the backing, and extend above the surface of the film as shown, typically a distance of 100 to 2000 μm above the film surface.

The grid lines are formed on the substrate by laying down an uncured or otherwise flowable resin or elastomer solution in an array grid, allowing the material to infiltrate the porous film down to the backing, then curing or otherwise hardening the grid lines to form the cell-array substrate.

One preferred material for the grid is a flowable silicone available from Loctite Corporation. The barrier material can be extruded through a narrow syringe (e.g., 22 gauge) using air pressure or mechanical pressure. The syringe is moved relative to the solid support to print the barrier elements as a grid pattern. The extruded bead of silicone wicks into the pores of the solid support and cures to form a shallow waterproof barrier separating the regions of the solid support.

In alternative embodiments, the barrier element can be a wax-based material or a thermoset material such as epoxy. The barrier material can also be a UV-curing polymer which is exposed to UV light after being printed onto the solid support. The barrier material may also be applied to the solid support using printing techniques such as silk-screen printing. The barrier material may also be a heat-seal stamping of the porous solid support which seals its pores and forms a water-impervious barrier element. The barrier material may also be a shallow grid which is laminated or otherwise adhered to the solid support.

In addition to plastic-backed nitrocellulose, the solid support can be virtually any porous membrane with or without a non-porous backing. Such membranes are readily available from numerous vendors and are made from nylon, PVDF, polysulfone and the like. In an alternative embodiment, the barrier element may also be used to adhere the porous membrane to a non-porous backing in addition to functioning as a barrier to prevent cross contamination of the assay reagents.

In an alternative embodiment, the solid support can be of a non-porous material. The barrier can be printed either before or after the microarray of biomolecules is printed on the solid support.

As can be appreciated, the cells formed by the grid lines and the underlying backing are water-impermeable, having side barriers projecting above the porous film in the cells. Thus, defined-volume samples can be placed in each well without risk of cross-contamination with sample material in adjacent cells. In Fig. 11, defined volume samples, such as sample 134, are shown in the cells.

As noted above, each well contains a microarray of distinct biopolymers. In one general embodiment, the

microarrays in the well are identical arrays of distinct biopolymers, e.g., different sequence polynucleotides. Such arrays can be formed in accordance with the methods described in Section II, by
5 depositing a first selected polynucleotide at the same selected microarray position in each of the cells, then depositing a second polynucleotide at a different microarray position in each well, and so on until a complete, identical microarray is formed in each cell.

10 In a preferred embodiment, each microarray contains about 10^3 distinct polynucleotide or polypeptide biopolymers per surface area of less than about 1 cm^2 . Also in a preferred embodiment, the biopolymers in each microarray region are present in a
15 defined amount between about 0.1 femtomoles and 100 nanomoles. The ability to form high-density arrays of biopolymers, where each region is formed of a well-defined amount of deposited material, can be achieved in accordance with the microarray-forming method
20 described in Section II.

Also in a preferred embodiments, the biopolymers are polynucleotides having lengths of at least about 50 bp, i.e., substantially longer than oligonucleotides which can be formed in high-density arrays by schemes
25 involving parallel, step-wise polymer synthesis on the array surface.

In the case of a polynucleotide array, in an assay procedure, a small volume of the labeled DNA probe mixture in a standard hybridization solution is loaded
30 onto each cell. The solution will spread to cover the entire microarray and stop at the barrier elements. The solid support is then incubated in a humid chamber at the appropriate temperature as required by the assay.

Each assay may be conducted in an "open-face" format where no further sealing step is required, since the hybridization solution will be kept properly hydrated by the water vapor in the humid chamber. At the conclusion of the incubation step, the entire solid support containing the numerous microarrays is rinsed quickly enough to dilute the assay reagents so that no significant cross contamination occurs. The entire solid support is then reacted with detection reagents if needed and analyzed using standard colorimetric, radioactive or fluorescent detection means. All processing and detection steps are performed simultaneously to all of the microarrays on the solid support ensuring uniform assay conditions for all of the microarrays on the solid support.

B. Glass-Slide Polynucleotide Array

Fig. 5 shows a substrate 136 formed according to another aspect of the invention, and intended for use in detecting binding of labeled polynucleotides to one or more of a plurality distinct polynucleotides. The substrate includes a glass substrate 138 having formed on its surface, a coating of a polycationic polymer, preferably a cationic polypeptide, such as polylysine or polyarginine. Formed on the polycationic coating is a microarray 140 of distinct polynucleotides, each localized at known selected array regions, such as regions 142.

The slide is coated by placing a uniform-thickness film of a polycationic polymer, e.g., poly-l-lysine, on the surface of a slide and drying the film to form a dried coating. The amount of polycationic polymer added is sufficient to form at least a monolayer of polymers on the glass surface. The polymer film is bound to surface via electrostatic binding between

negative silyl-OH groups on the surface and charged amine groups in the polymers. Poly-L-lysine coated glass slides may be obtained commercially, e.g., from Sigma Chemical Co. (St. Louis, MO).

5 To form the microarray, defined volumes of distinct polynucleotides are deposited on the polymer-coated slide, as described in Section II. According to an important feature of the substrate, the deposited polynucleotides remain bound to the coated slide
10 surface non-covalently when an aqueous DNA sample is applied to the substrate under conditions which allow hybridization of reporter-labeled polynucleotides in the sample to complementary-sequence (single-stranded) polynucleotides in the substrate array. The method is
15 illustrated in Examples 1 and 2.

 To illustrate this feature, a substrate of the type just described, but having an array of same-sequence polynucleotides, was mixed with fluorescent-labeled complementary DNA under hybridization
20 conditions. After washing to remove non-hybridized material, the substrate was examined by low-power fluorescence microscopy. The array can be visualized by the relatively uniform labeling pattern of the array regions.

25 In a preferred embodiment, each microarray contains at least 10^3 distinct polynucleotide or polypeptide biopolymers per surface area of less than about 1 cm^2 . In the embodiment shown in Fig. 5, the microarray contains 400 regions in an area of about 16
30 mm^2 , or 2.5×10^3 regions/ cm^2 . Also in a preferred embodiment, the polynucleotides in the each microarray region are present in a defined amount between about 0.1 femtomoles and 100 nanomoles in the case of polynucleotides. As above, the ability to form high-

density arrays of this type, where each region is formed of a well-defined amount of deposited material, can be achieved in accordance with the microarray-forming method described in Section II.

5 Also in a preferred embodiments, the polynucleotides have lengths of at least about 50 bp, i.e., substantially longer than oligonucleotides which can be formed in high-density arrays by various in situ synthesis schemes.

10

V. Utility

Microarrays of immobilized nucleic acid sequences prepared in accordance with the invention can be used for large scale hybridization assays in numerous
15 genetic applications, including genetic and physical mapping of genomes, monitoring of gene expression, DNA sequencing, genetic diagnosis, genotyping of organisms, and distribution of DNA reagents to researchers.

For gene mapping, a gene or a cloned DNA fragment
20 is hybridized to an ordered array of DNA fragments, and the identity of the DNA elements applied to the array is unambiguously established by the pixel or pattern of pixels of the array that are detected. One application of such arrays for creating a genetic map is described
25 by Nelson, et al. (1993). In constructing physical maps of the genome, arrays of immobilized cloned DNA fragments are hybridized with other cloned DNA fragments to establish whether the cloned fragments in the probe mixture overlap and are therefore contiguous
30 to the immobilized clones on the array. For example, Lehrach, et al., describe such a process.

The arrays of immobilized DNA fragments may also be used for genetic diagnostics. To illustrate, an array containing multiple forms of a mutated gene or
35 genes can be probed with a labeled mixture of a

patient's DNA which will preferentially interact with only one of the immobilized versions of the gene.

The detection of this interaction can lead to a medical diagnosis. Arrays of immobilized DNA fragments can also be used in DNA probe diagnostics. For example, the identity of a pathogenic microorganism can be established unambiguously by hybridizing a sample of the unknown pathogen's DNA to an array containing many types of known pathogenic DNA. A similar technique can also be used for unambiguous genotyping of any organism. Other molecules of genetic interest, such as cDNA's and RNA's can be immobilized on the array or alternately used as the labeled probe mixture that is applied to the array.

In one application, an array of cDNA clones representing genes is hybridized with total cDNA from an organism to monitor gene expression for research or diagnostic purposes. Labeling total cDNA from a normal cell with one color fluorophore and total cDNA from a diseased cell with another color fluorophore and simultaneously hybridizing the two cDNA samples to the same array of cDNA clones allows for differential gene expression to be measured as the ratio of the two fluorophore intensities. This two-color experiment can be used to monitor gene expression in different tissue types, disease states, response to drugs, or response to environmental factors. & An example of this approach is illustrated in Examples 2, described with respect to Fig. 8.

By way of example and without implying a limitation of scope, such a procedure could be used to simultaneously screen many patients against all known mutations in a disease gene. This invention could be used in the form of, for example, 96 identical 0.9 cm x 2.2 cm microarrays fabricated on a single 12 cm x 18 cm

sheet of plastic-backed nitrocellulose where each microarray could contain, for example, 100 DNA fragments representing all known mutations of a given gene. The region of interest from each of the DNA samples from 96 patients could be amplified, labeled, and hybridized to the 96 individual arrays with each assay performed in 100 microliters of hybridization solution. The approximately 1 thick silicone rubber barrier elements between individual arrays prevent cross contamination of the patient samples by sealing the pores of the nitrocellulose and by acting as a physical barrier between each microarray. The solid support containing all 96 microarrays assayed with the 96 patient samples is incubated, rinsed, detected and analyzed as a single sheet of material using standard radioactive, fluorescent, or colorimetric detection means (Maniatis, et al., 1989). Previously, such a procedure would involve the handling, processing and tracking of 96 separate membranes in 96 separate sealed chambers. By processing all 96 arrays as a single sheet of material, significant time and cost savings are possible.

The assay format can be reversed where the patient or organism's DNA is immobilized as the array elements and each array is hybridized with a different mutated allele or genetic marker. The gridded solid support can also be used for parallel non-DNA ELISA assays. Furthermore, the invention allows for the use of all standard detection methods without the need to remove the shallow barrier elements to carry out the detection step.

In addition to the genetic applications listed above, arrays of whole cells, peptides, enzymes, antibodies, antigens, receptors, ligands, phospholipids, polymers, drug cogener preparations or

chemical substances can be fabricated by the means described in this invention for large scale screening assays in medical diagnostics, drug discovery, molecular biology, immunology and toxicology.

5 The multi-cell substrate aspect of the invention allows for the rapid and convenient screening of many DNA probes against many ordered arrays of DNA fragments. This eliminates the need to handle and detect many individual arrays for performing mass
10 screenings for genetic research and diagnostic applications. Numerous microarrays can be fabricated on the same solid support and each microarray reacted with a different DNA probe while the solid support is processed as a single sheet of material.

15

The following examples illustrate, but in no way are intended to limit, the present invention.

Example 1

20 Genomic-Complexity Hybridization to Micro
DNA Arrays Representing the Yeast
Saccharomyces cerevisiae Genome with
Two-Color Fluorescent Detection

The array elements were randomly amplified PCR
25 (Bohlander, et al., 1992) products using physically mapped lambda clones of *S. cerevisiae* genomic DNA templates (Riles, et al., 1993). The PCR was performed directly on the lambda phage lysates resulting in an amplification of both the 35 kb lambda vector and the
30 5-15 kb yeast insert sequences in the form of a uniform distribution of PCR product between 250-1500 base pairs in length. The PCR product was purified using Sephadex G50 gel filtration (Pharmacia, Piscataway, NJ) and concentrated by evaporation to dryness at room
35 temperature overnight. Each of the 864 amplified

lambda clones was rehydrated in 15 μ l of 3 \times SSC in preparation for spotting onto the glass.

The micro arrays were fabricated on microscope slides which were coated with a layer of poly-l-lysine (Sigma). The automated apparatus described in Section IV loaded 1 μ l of the concentrated lambda clone PCR product in 3 \times SSC directly from 96 well storage plates into the open capillary printing element and deposited ~5 nl of sample per slide at 380 micron spacing between spots, on each of 40 slides. The process was repeated for all 864 samples and 8 control spots. After the spotting operation was complete, the slides were rehydrated in a humid chamber for 2 hours, baked in a dry 80° vacuum oven for 2 hours, rinsed to remove unabsorbed DNA and then treated with succinic anhydride to reduce non-specific adsorption of the labeled hybridization probe to the poly-l-lysine coated glass surface. Immediately prior to use, the immobilized DNA on the array was denatured in distilled water at 90° for 2 minutes.

For the pooled chromosome experiment, the 16 chromosomes of *Saccharomyces cerevisiae* were separated in a CHEF agarose gel apparatus (Biorad, Richmond, CA). The six largest chromosomes were isolated in one gel slice and the smallest 10 chromosomes in a second gel slice. The DNA was recovered using a gel extraction kit (Qiagen, Chatsworth, CA). The two chromosome pools were randomly amplified in a manner similar to that used for the target lambda clones. Following amplification, 5 micrograms of each of the amplified chromosome pools were separately random-primer labeled using Klenow polymerase (Amersham, Arlington Heights, IL) with a lissamine conjugated nucleotide analog (Dupont NEN, Boston, MA) for the pool containing the six largest chromosomes, and with a fluorescein

conjugated nucleotide analog (BMB) for the pool containing smallest ten chromosomes. The two pools were mixed and concentrated using an ultrafiltration device (Amicon, Danvers, MA).

5 Five micrograms of the hybridization probe consisting of both chromosome pools in 7.5 μ l of TE was denatured in a boiling water bath and then snap cooled on ice. 2.5 μ l of concentrated hybridization solution (5 \times SSC and 0.1% SDS) was added and all 10 μ l
10 transferred to the array surface, covered with a cover slip, placed in a custom-built single-slide humidity chamber and incubated at 60° for 12 hours. The slides were then rinsed at room temperature in 0.1 \times SSC and 0.1%SDS for 5 minutes, cover slipped and scanned.

15 A custom built laser fluorescent scanner was used to detect the two-color hybridization signals from the 1.8 \times 1.8 cm array at 20 micron resolution. The scanned image was gridded and analyzed using custom image analysis software. After correcting for optical
20 crosstalk between the fluorophores due to their overlapping emission spectra, the red and green hybridization values for each clone on the array were correlated to the known physical map position of the clone resulting in a computer-generated color karyotype
25 of the yeast genome.

Figure 6 shows the hybridization pattern of the two chromosome pools. A red signal indicates that the lambda clone on the array surface contains a cloned genomic DNA segment from one of the largest six yeast
30 chromosomes. A green signal indicates that the lambda clone insert comes from one of the smallest ten yeast chromosomes. Orange signals indicate repetitive sequences which cross hybridized to both chromosome pools. Control spots on the array confirm that the
35 hybridization is specific and reproducible.

The physical map locations of the genomic DNA fragments contained in each of the clones used as array elements have been previously determined by Olson and co-workers (Riles, et al.) allowing for the automatic generation of the color karyotype shown in Figure 7. The color of a chromosomal section on the karyotype corresponds to the color of the array element containing the clone from that section. The black regions of the karyotype represent false negative dark spots on the array (10%) or regions of the genome not covered by the Olson clone library (90%). Note that the largest six chromosomes are mainly red while the smallest ten chromosomes are mainly green matching the original CHEF gel isolation of the hybridization probe. Areas of the red chromosomes containing green spots and vice-versa are probably due to spurious sample tracking errors in the formation of the original library and in the amplification and spotting procedures.

The yeast genome arrays have also been probed with individual clones or pools of clones that are fluorescently labeled for physical mapping purposes. The hybridization signals of these clones to the array were translated into a position on the physical map of yeast.

25

Example 2

Total cDNA Hybridized to Micro Arrays of cDNA Clones with Two-Color Fluorescent Detection

24 clones containing cDNA inserts from the plant *Arabidopsis* were amplified using PCR. Salt was added to the purified PCR products to a final concentration of 3 × SSC. The cDNA clones were spotted on poly-l-lysine coated microscope slides in a manner similar to Example 1. Among the cDNA clones was a clone

35

representing a transcription factor HAT 4, which had previously been used to create a transgenic line of the plant *Arabidopsis*, in which this gene is present at ten times the level found in wild-type *Arabidopsis* (Schena, et al., 1992).

Total poly-A mRNA from wild type *Arabidopsis* was isolated using standard methods (Maniatis, et al., 1989) and reverse transcribed into total cDNA, using fluorescein nucleotide analog to label the cDNA product (green fluorescence). A similar procedure was performed with the transgenic line of *Arabidopsis* where the transcription factor HAT4 was inserted into the genome using standard gene transfer protocols. cDNA copies of mRNA from the transgenic plant are labeled with a lissamine nucleotide analog (red fluorescence). Two micrograms of the cDNA products from each type of plant were pooled together and hybridized to the cDNA clone array in a 10 microliter hybridization reaction in a manner similar to Example 1. Rinsing and detection of hybridization was also performed in a manner similar to Example 1. Fig. 8 show the resulting hybridization pattern of the array.

Genes equally expressed in wild type and the transgenic *Arabidopsis* appeared yellow due to equal contributions of the green and red fluorescence to the final signal. The dots are different intensities of yellow indicating various levels of gene expression. The cDNA clone representing the transcription factor HAT4, expressed in the transgenic line of *Arabidopsis* but not detectably expressed in wild type *Arabidopsis*, appears as a red dot (with the arrow pointing to it), indicating the preferential expression of the transcription factor in the red-labeled transgenic *Arabidopsis* and the relative lack of expression of the

transcription factor in the green-labeled wild type *Arabidopsis*.

An advantage of the microarray hybridization format for gene expression studies is the high partial concentration of each cDNA species achievable in the 10 microliter hybridization reaction. This high partial concentration allows for detection of rare transcripts without the need for PCR amplification of the hybridization probe which may bias the true genetic representation of each discrete cDNA species.

Gene expression studies such as these can be used for genomics research to discover which genes are expressed in which cell types, disease states, development states or environmental conditions. Gene expression studies can also be used for diagnosis of disease by empirically correlating gene expression patterns to disease states.

Example 3

Multiplexed Colorimetric Hybridization on a Gridded Solid Support

A sheet of plastic-backed nitrocellulose was gridded with barrier elements made from silicone rubber according to the description in Section IV-A. The sheet was soaked in 10 × SSC and allowed to dry. As shown in Fig. 12, 192 M13 clones each with a different yeast inserts were arrayed 400 microns apart in four quadrants of the solid support using the automated device described in Section III. The bottom left quadrant served as a negative control for hybridization while each of the other three quadrants was hybridized simultaneously with a different oligonucleotide using the open-face hybridization technology described in Section IV-A. The first two and last four elements of

each array are positive controls for the colorimetric detection step.

The oligonucleotides were labeled with fluorescein which was detected using an anti-fluorescein antibody
5 conjugated to alkaline phosphatase that precipitated an NBT/BCIP dye on the solid support (Amersham). Perfect matches between the labeled oligos and the M13 clones resulted in dark spots visible to the naked eye and detected using an optical scanner (HP ScanJet II)
10 attached to a personal computer. The hybridization patterns are different in every quadrant indicating that each oligo found several unique M13 clones from among the 192 with a perfect sequence match. Note that the open capillary printing tip leaves detectable
15 dimples on the nitrocellulose which can be used to automatically align and analyze the images.

Although the invention has been described with respect to specific embodiments and methods, it will be
20 clear that various changes and modification may be made without departing from the invention.

IT IS CLAIMED:

1. A method of forming a microarray of analyte-assay regions on a solid support, where each region in
5 the array has a known amount of a selected, analyte-specific reagent, said method comprising,
 - (a) loading a solution of a selected analyte-specific reagent in a reagent-dispensing device having an elongate capillary channel (i) formed by spaced-
10 apart, coextensive elongate members, (ii) adapted to hold a quantity of the reagent solution and (iii) having a tip region at which aqueous solution in the channel forms a meniscus,
 - (b) tapping the tip of the dispensing device
15 against a solid support at a defined position on the surface, with an impulse effective to break the meniscus in the capillary channel and deposit a selected volume of solution on the surface, and
 - (c) repeating steps (a) and (b) until said array
20 is formed.
2. The method of claim 1, wherein said tapping is carried out with an impulse effective to deposit a
25 selected volume in the volume range between 0.01 to 100 nl.
3. The method of claim 1, wherein said channel is formed by a pair of spaced-apart tapered elements.
- 30 4. The method of claim 1, for forming a plurality of such arrays, wherein step (b) is applied to a selected position on each of a plurality of solid supports at each repeat cycle proceeding step (c).

5. The method of claim 1, which further includes, after performing steps (a) and (b) at least one time, reloading the reagent-dispensing device with a new reagent solution by the steps of (i) dipping the
5 capillary channel of the device in a wash solution, (ii) removing wash solution drawn into the capillary channel, and (iii) dipping the capillary channel into the new reagent solution.

10 6. Automated apparatus for forming a microarray of analyte-assay regions on a plurality of solid supports, where each region in the array has a known amount of a selected, analyte-specific reagent, said apparatus comprising

15 (a) a holder for holding, at known positions, a plurality of planar supports,

(b) a reagent dispensing device having an open capillary channel (i) formed by spaced-apart, coextensive elongate members (ii) adapted to hold a
20 quantity of the reagent solution and (iii) having a tip region at which aqueous solution in the channel forms a meniscus,

(c) positioning means for positioning the dispensing device at a selected array position with
25 respect to a support in said holder,

(d) dispensing means for moving the device into tapping engagement against a support with a selected impulse, when the device is positioned at a defined array position with respect to that support, with an
30 impulse effective to break the meniscus of liquid in the capillary channel and deposit a selected volume of solution on the surface, and

(e) control means for controlling said positioning and dispensing means.

7. The apparatus of claim 6, wherein said dispensing means is effective to move said dispensing device against a support with an impulse effective to deposit a selected volume in the volume range between
5 0.01 to 100 nl.

8. The apparatus of claim 6, wherein said channel is formed by a pair of spaced-apart tapered elements.

10 9. The apparatus of claim 6, wherein the control means operates to (i) place the dispensing device at a loading station, (ii) move the capillary channel in the device into a selected reagent at the loading station, to load the dispensing device with the reagent, and
15 (iii) dispense the reagent at a defined array position on each of the supports on said holder.

10. The apparatus of claim 6, wherein the control device further operates, at the end of a dispensing
20 cycle, to wash the dispensing device by (i) placing the dispensing device at a washing station, (ii) moving the capillary channel in the device into a wash fluid, to load the dispensing device with the fluid, and (iii) remove the wash fluid prior to loading the dispensing
25 device with a fresh selected reagent.

11. The apparatus of claim 6, wherein said device is one of a plurality of such devices which are carried on the arm for dispensing different analyte assay
30 reagents at selected spaced array positions.

12. A substrate with a surface having a microarray of at least 10^3 distinct polynucleotide or polypeptide biopolymers per 1 cm^2 surface area, each

distinct biopolymer sample (i) being disposed at a separate, defined position in said array, (ii) having a length of at least 50 subunits, and (iii) being present in a defined amount between about 0.1 femtomole and 100 nanomoles.

13. The substrate of claim 12, wherein said surface is glass slide coated with polylysine, and said biopolymers are polynucleotides.

10

14. The substrate of claim 12, wherein said substrate has a water-impermeable backing, a water-permeable film formed on the backing, and a grid formed on the film, where said grid (i) is composed of intersecting water-impervious grid elements extending from said backing to positions raised above the surface of said film, and (ii) partitions the film into a plurality of water-impervious cells, where each cell contains such a biopolymer array.

20

15. A substrate with a surface array of sample-receiving cells, comprising
a water-impermeable backing,
a water-permeable film formed on the backing, and
a grid formed on the film, said grid being composed of intersecting water-impervious grid elements extending from said backing to positions raised above the surface of said film.

16. The substrate of claim 15, wherein the cells of the array each contain an array of biopolymers.

17. A substrate for use in detecting binding of labeled biopolymers to one or more of a plurality distinct polynucleotides, comprising

35

a non-porous, glass substrate,
a coating of a cationic polymer on said substrate,
and

an array of distinct polynucleotides to said
5 coating, where each biopolymer is disposed at a
separate, defined position in a surface array of
biopolymers.

18. A method of detecting differential expression
10 of each of a plurality of genes in a first cell type
with respect to expression of the same genes in a
second cell types, said method comprising

producing fluorescence-labeled cDNA's from mRNA's
isolated from the two cells types, where the cDNA's
15 from the first and second cells are labeled with first
and second different fluorescent reporters,

adding a mixture of the labeled cDNA's from the
two cell types to an array of polynucleotides
representing a plurality of known genes derived from
20 the two cell types, under conditions that result in
hybridization of the cDNA's to complementary-sequence
polynucleotides in the array; and

examining the array by fluorescence under
fluorescence excitation conditions in which (i)
25 polynucleotides in the array that are hybridized
predominantly to cDNA's derived from one of the first
and second cell types give a distinct first or second
fluorescence emission color, respectively, and (ii)
polynucleotides in the array that are hybridized to
30 substantially equal numbers of cDNA's derived from the
first and second cell types give a distinct combined
fluorescence emission color, respectively,

wherein the relative expression of known genes in
the two cell types can be determined by the observed
35 fluorescence emission color of each spot.

19. The method of claim 18, wherein the array of polynucleotides is formed on a substrate with a surface having an array of at least 10^2 distinct polynucleotide or polypeptide biopolymers in a surface area of less than about 1 cm^2 , each distinct biopolymer (i) being
5 disposed at a separate, defined position in said array, (ii) having a length of at least 50 subunits, and (iii) being present in a defined amount between about .1 femtomole and 100 nmoles.

10

20. The method of claim 19, wherein said surface is a glass slide coated with polylysine, and said biopolymers are polynucleotides non-covalently bound to said polylysine.

15

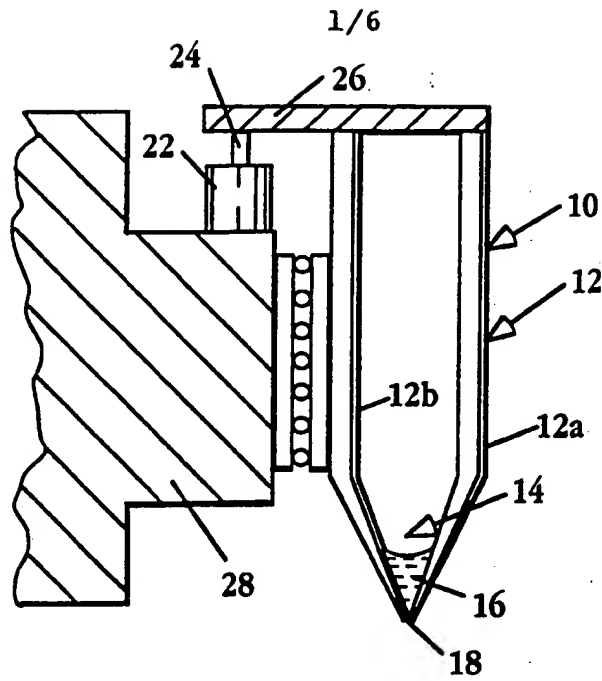


Fig. 1

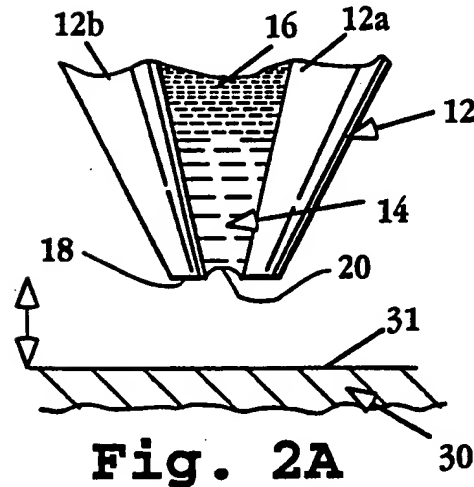


Fig. 2A

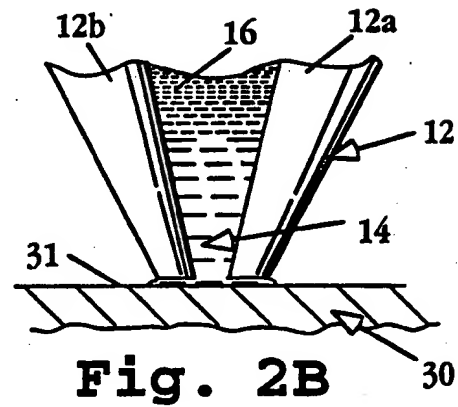


Fig. 2B

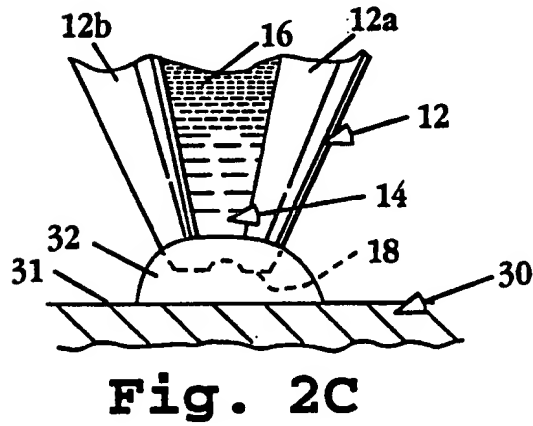
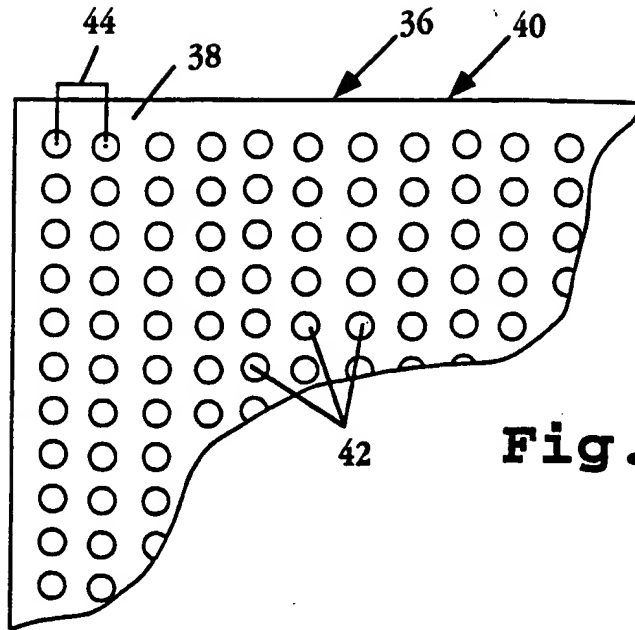
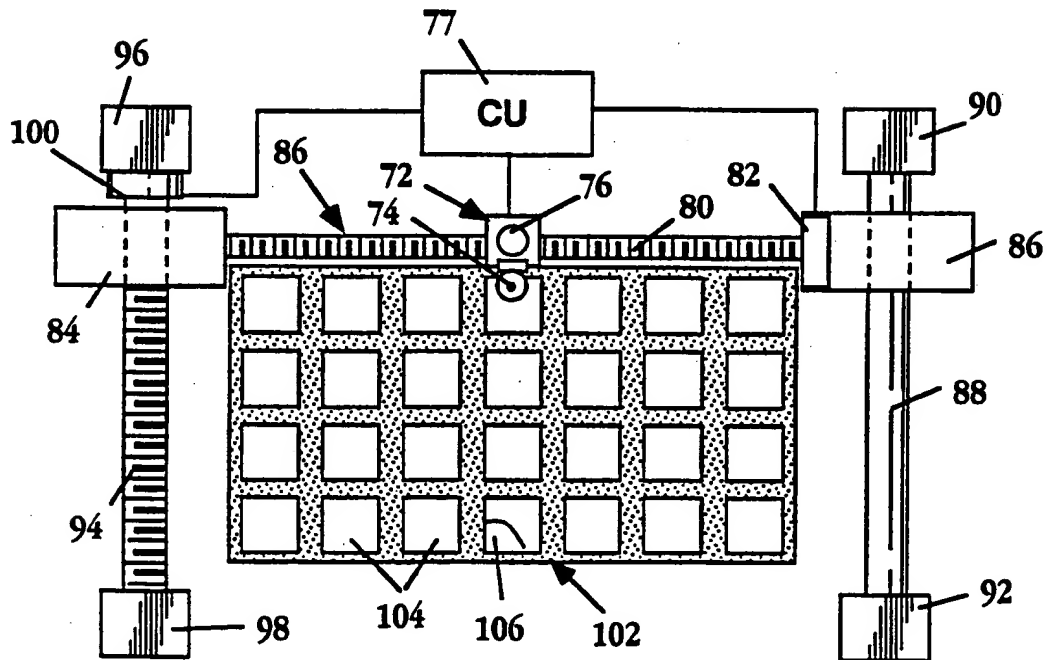


Fig. 2C

2/6

**Fig. 3****Fig. 4**

3/6

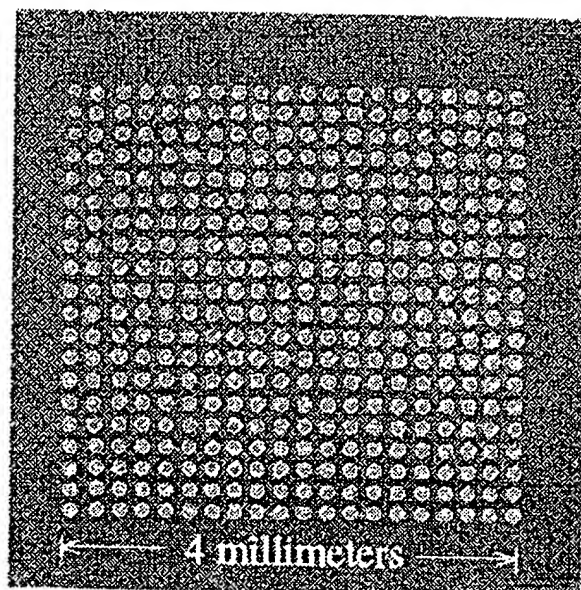


Fig. 5

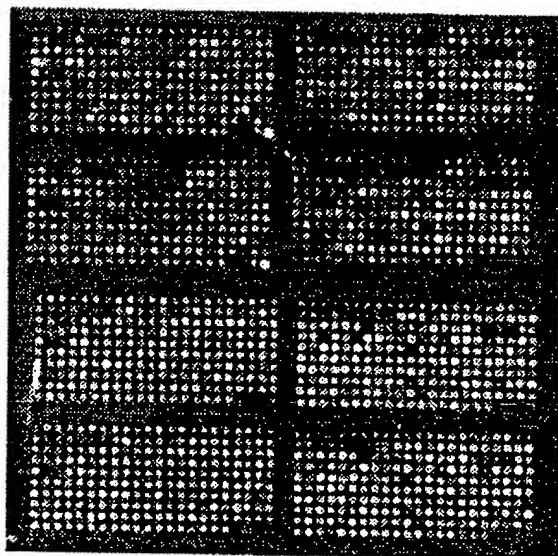


Fig. 6

SUBSTITUTE SHEET (RULE 26)

4/6

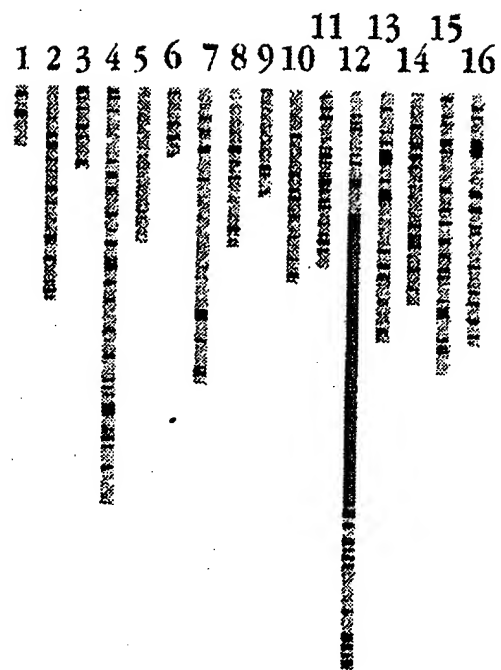


Fig. 7

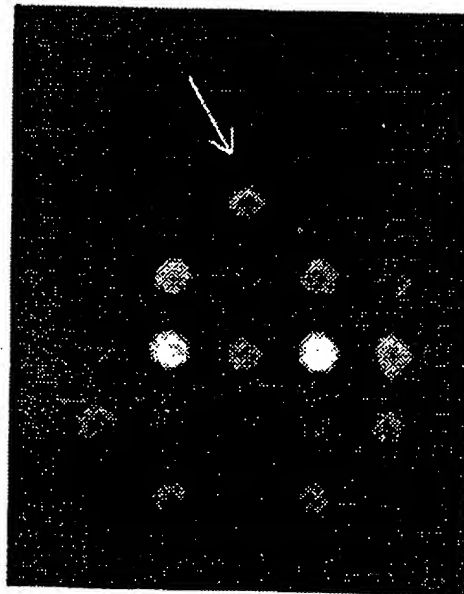


Fig. 8

SUBSTITUTE SHEET (RULE 26)

5/6

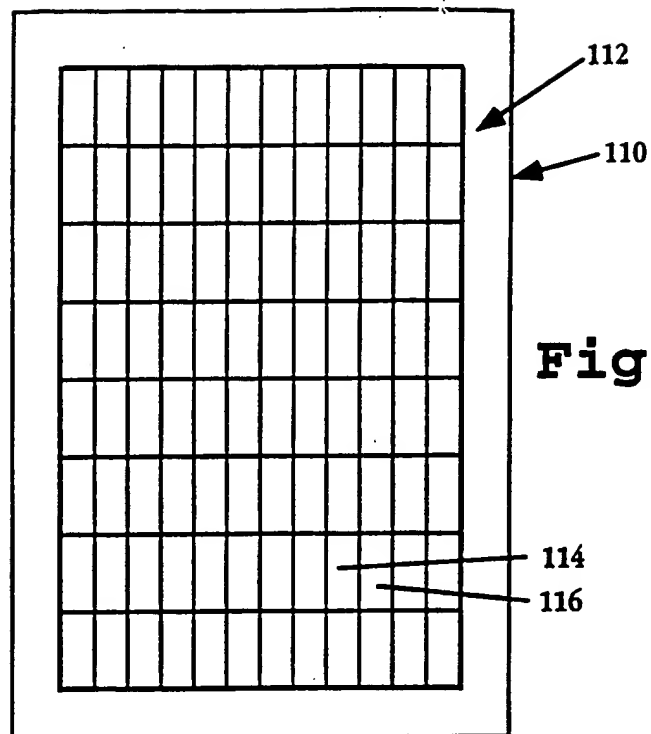


Fig. 9

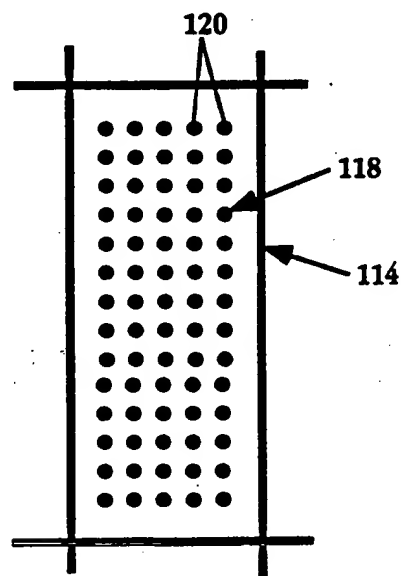
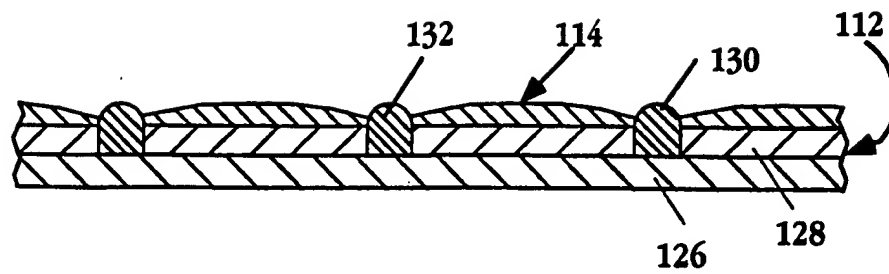
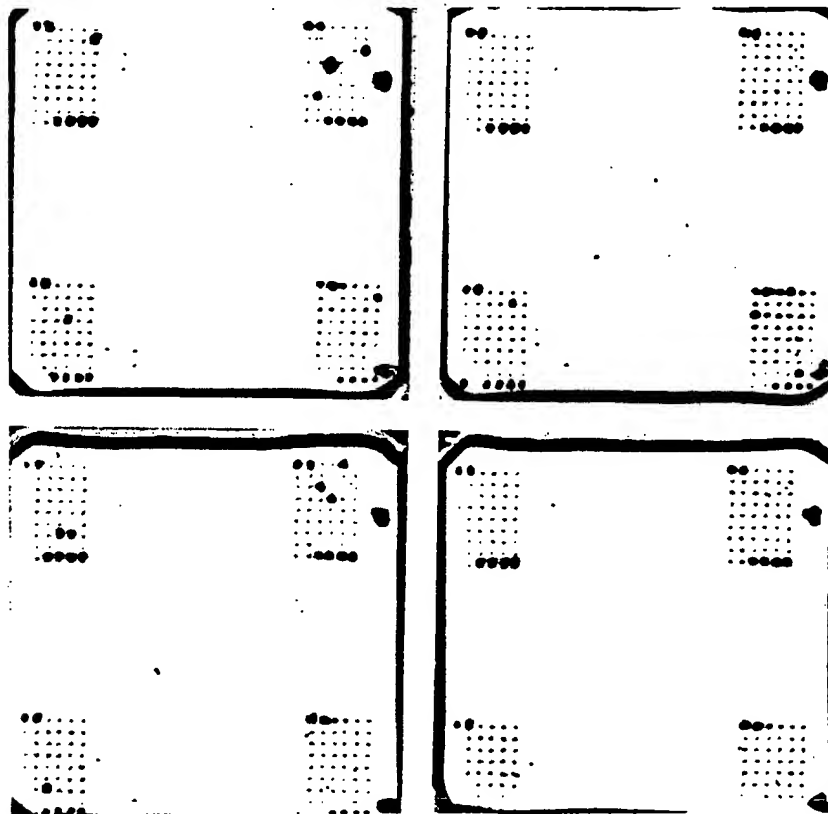


Fig. 10

6/6

**Fig. 11****Fig. 12**

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US95/07659

A. CLASSIFICATION OF SUBJECT MATTER

IPC(6) : G01N 33/543, 33/68

US CL : 435/6; 436/518

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 422/57; 435/4,6,973; 436/518,524,527,531,805,809

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A,P	US, A, 5,338,688 (DEEG ET AL) 16 August 1994, see entire document	1-17
A	US, A, 5,204,268 (MATSUMOTO) 20 April 1993, see entire document.	6-11
A	US, A, 4,071,315 (CHATEAU) 31 January 1978, see entire document.	12-17
A	US, A, 5,100,777 (CHANG) 31 March 1992, see entire document.	12-17
A	US, A, 5,200,312 (OPRANDY) 06 April 1993, see entire document.	12-17

☐ Further documents are listed in the continuation of Box C.
 ☐ See patent family annex.

* Special categories of cited documents:	T	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
A document defining the general state of the art which is not considered to be of particular relevance	X*	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
E earlier document published on or after the international filing date	Y*	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
L document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	&*	document member of the same patent family
O document referring to an oral disclosure, use, exhibition or other means		
P document published prior to the international filing date but later than the priority date claimed		

Date of the actual completion of the international search

15 SEPTEMBER 1995

Date of mailing of the international search report

06 OCT 1995

Name and mailing address of the ISA/US
Commissioner of Patents and Trademarks
Box PCT
Washington, D.C. 20231

Facsimile No. (703) 305-3230

Authorized officer
CHRISTOPHER CHIN

Telephone No. (703) 308-0196

Discovery and analysis of inflammatory disease-related genes using cDNA microarrays

(inflammation/human genome analysis/gene discovery)

RENU A. HELLER*[†], MARK SCHENA*, ANDREW CHAI*, DARI SHALON[‡], TOD BEDILION[‡], JAMES GILMORE[‡], DAVID E. WOOLLEY[§], AND RONALD W. DAVIS*

*Department of Biochemistry, Beckman Center, Stanford University Medical Center, Stanford, CA 94305; [‡]Synteni, Palo Alto, CA 94306; and [§]Department of Medicine, Manchester Royal Infirmary, Manchester, United Kingdom

Contributed by Ronald W. Davis, December 27, 1996

ABSTRACT cDNA microarray technology is used to profile complex diseases and discover novel disease-related genes. In inflammatory disease such as rheumatoid arthritis, expression patterns of diverse cell types contribute to the pathology. We have monitored gene expression in this disease state with a microarray of selected human genes of probable significance in inflammation as well as with genes expressed in peripheral human blood cells. Messenger RNA from cultured macrophages, chondrocyte cell lines, primary chondrocytes, and synoviocytes provided expression profiles for the selected cytokines, chemokines, DNA binding proteins, and matrix-degrading metalloproteinases. Comparisons between tissue samples of rheumatoid arthritis and inflammatory bowel disease verified the involvement of many genes and revealed novel participation of the cytokine interleukin 3, chemokine Gro α and the metalloproteinase matrix metallo-elastase in both diseases. From the peripheral blood library, tissue inhibitor of metalloproteinase 1, ferritin light chain, and manganese superoxide dismutase genes were identified as expressed differentially in rheumatoid arthritis compared with inflammatory bowel disease. These results successfully demonstrate the use of the cDNA microarray system as a general approach for dissecting human diseases.

The recently described cDNA microarray or DNA-chip technology allows expression monitoring of hundreds and thousands of genes simultaneously and provides a format for identifying genes as well as changes in their activity (1, 2). Using this technology, two-color fluorescence patterns of differential gene expression in the root versus the shoot tissue of *Arabidopsis* were obtained in a specific array of 48 genes (1). In another study using a 1000 gene array from a human peripheral blood library, novel genes expressed by T cells were identified upon heat shock and protein kinase C activation (3).

The technology uses cDNA sequences or cDNA inserts of a library for PCR amplification that are arrayed on a glass slide with high speed robotics at a density of 1000 cDNA sequences per cm². These microarrays serve as gene targets for hybridization to cDNA probes prepared from RNA samples of cells or tissues. A two-color fluorescence labeling technique is used in the preparation of the cDNA probes such that a simultaneous hybridization but separate detection of signals provides the comparative analysis and the relative abundance of specific genes expressed (1, 2). Microarrays can be constructed from specific cDNA clones of interest, a cDNA library, or a select number of open reading frames from a genome sequencing database to allow a large-scale functional analysis of expressed sequences.

Because of the wide spectrum of genes and endogenous mediators involved, the microarray technology is well suited for analyzing chronic diseases. In rheumatoid arthritis (RA), inflammation of the joint is caused by the gene products of many different cell types present in the synovium and cartilage tissues plus those infiltrating from the circulating blood. The autoimmune and inflammatory nature of the disease is a cumulative result of genetic susceptibility factors and multiple responses, paracrine and autocrine in nature, from macrophages, T cells, plasma cells, neutrophils, synovial fibroblasts, chondrocytes, etc. Growth factors, inflammatory cytokines (4), and the chemokines (5) are the important mediators of this inflammatory process. The ensuing destruction of the cartilage and bone by the invading synovial tissue includes the actions of prostaglandins and leukotrienes (6), and the matrix degrading metalloproteinases (MMPs). The MMPs are an important class of Zn-dependent metallo-endoproteases that can collectively degrade the proteoglycan and collagen components of the connective tissue matrix (7).

This paper presents a study in which the involvement of select classes of molecules in RA was examined. Also investigated were 1000 human genes randomly selected from a peripheral human blood cell library. Their differential and quantitative expression analysis in cells of the joint tissue, in diseased RA tissue and in inflammatory bowel disease (IBD) tissues was conducted to demonstrate the utility of the microarray method to analyze complex diseases by their pattern of gene expression. Such a survey provides insight not only into the underlying cause of the pathology, but also provides the opportunity to selectively target genes for disease intervention by appropriate drug development and gene therapies.

METHODS

Microarray Design, Development, and Preparation. Two approaches for the fabrication of cDNA microarrays were used in this study. In the first approach, known human genes of probable significance in RA were identified. Regions of the clones, preferably 1 kb in length, were selected by their proximity to the 3' end of the cDNA and for areas of least identity to related and repetitive sequences. Primers were synthesized to amplify the target regions by standard PCR protocols (3). Products were

Abbreviations: RA, rheumatoid arthritis; MMP, matrix-degrading metalloproteinase; IBD, inflammatory bowel disease; LPS, lipopolysaccharide; PMA, phorbol 12-myristate 13-acetate; TNF- α , tumor necrosis factor α ; IL, interleukin; TGF- β , transforming growth factor β ; GCSF, granulocyte colony-stimulating factor; MIP, macrophage inflammatory protein; MIF, migration inhibitory factor; HME, human matrix metallo-elastase; RANTES, regulated upon activation, normal T cell expressed and secreted; Gel, gelatinase; VCAM, vascular cell adhesion molecule; ICE, IL-1 converting enzyme; PUMP, putative metalloproteinase; MnSOD, manganese superoxide dismutase; TIMP, tissue inhibitor of metalloproteinase; MCP, macrophage chemotactic protein.

[†]To whom reprint requests should be sent at the present address: Roche Bioscience, S3-1, 3401 Hillview Avenue, Palo Alto, CA 94304.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Copyright © 1997 by THE NATIONAL ACADEMY OF SCIENCES OF THE USA
0027-8424/97/942150-6\$2.00/0
PNAS is available online at <http://www.pnas.org>.

verified by gel electrophoresis and purified with Qiaquick 96-well purification kit (Qiagen, Chatsworth, CA), lyophilized (Savant), and resuspended in 5 μ l of 3 \times standard saline citrate (SSC) buffer for arraying. In the second approach, the microarray containing the 1056 human genes from the peripheral blood lymphocyte library was prepared as described (3).

Tissue Specimens. Rheumatoid synovial tissue was obtained from patients with late stage classic RA undergoing remedial synovectomy or arthroplasty of the knee. Synovial tissue was separated from any associated connective tissue or fat. One gram of each synovial specimen was subjected to RNA extraction within 40 min of surgical excision, or explants were cultured in serum-free medium to examine any changes under *in vitro* conditions. For IBD, specimens of macroscopically inflamed lower intestinal mucosa were obtained from patients with Crohn disease undergoing remedial surgery. The hypertrophied mucosal tissue was separated from underlying connective tissue and extracted for RNA.

Cultured Cells. The Mono Mac-6 (MM6) monocytic cells (8) were grown in RPMI medium. Human chondrosarcoma SW1353 cells, primary human chondrocytes, and synoviocytes (9, 10) were cultured in DMEM; all culture media were supplemented with 10% fetal bovine serum, 100 μ g/ml streptomycin, and 500 units/ml penicillin. Treatment of cells with lipopolysaccharide (LPS) endotoxin at 30 ng/ml, phorbol 12-myristate 13-acetate (PMA) at 50 ng/ml, tumor necrosis factor α (TNF- α) at 50 ng/ml, interleukin (IL)-1 β at 30 ng/ml, or transforming growth factor- β (TGF- β) at 100 ng/ml is described in the figure legends.

Fluorescent Probe, Hybridization, and Scanning. Isolation of mRNA, probe preparation, and quantitation with *Arabidopsis* control mRNAs was essentially as described (3) except for the following minor modification. Following the reverse transcriptase step, the appropriate Cy3- and Cy5-labeled samples were pooled; mRNA degraded by heating the sample to 65°C for 10 min with the addition of 5 μ l of 0.5M NaOH plus 0.5 ml of 10 mM EDTA. The pooled cDNA was purified from unincorporated nucleotides by gel filtration in Centri-spin columns (Princeton Separations, Adelphia, NJ). Samples were lyophilized and dissolved in 6 μ l of hybridization buffer (5 \times SSC plus 0.2% SDS). Hybridizations, washes, scanning, quantitation procedures, and pseudocolor representations of fluorescent images have been described (3). Scans for the two fluorescent probes were normalized either to the fluorescence intensity of *Arabidopsis* mRNAs spiked into the labeling reactions (see Figs. 2–4) or to the signal intensity of β -actin and glyceraldehyde-3-phosphate dehydrogenase (GAPDH; see Fig. 5).

RESULTS

Ninety-Six-Genes Microarray Design. The actions of cytokines, growth factors, chemokines, transcription factors, MMPs, prostaglandins, and leukotrienes are well recognized in inflammatory disease, particularly RA (11–14). Fig. 1 displays the selected genes for this study and also includes control cDNAs of housekeeping genes such as β -actin and GAPDH and genes from *Arabidopsis* for signal normalization and quantitation (row A, columns 1–12).

Defining Microarray Assay Conditions. Different lengths and concentrations of target DNA were tested by arraying PCR-

	1	2	3	4	5	6	7	8	9	10	11	12
A	BLANK	BLANK	HAT1 HAT1	HAT1 HAT1	HAT4 HAT4	HAT4 HAT4	HAT22 HAT22	HAT22 HAT22	YES23 YES23	YES23 YES23	BACTIN β -actin	G3PDH G3PDH
B	IL1A IL-1 α	IL1B IL-1 β	IL1RA IL-1RA	IL2 IL-2	IL3 IL-3	IL4 IL-4	IL6 IL-6	IL6R IL-6R	IL7 IL-7	CFOS c-fos	CJUN c-jun	RFRA1 Rat Fra-1
C	IL8 IL-8	IL9 IL-9	IL10 IL-10	ICE ICE	IFNG IFN γ	GCSF G-CSF	MCSF M-CSF	GMCSF GM-CSF	TNFB.1 TNF β	CREL c-rel	NFKB50 NF κ Bp50	NFKB65.1 NF κ Bp65
D	TNFA.1 TNF α	TNFA.2 TNF α	TNFA.3 TNF α	TNFA.4 TNF α	TNFA.5 TNF α	TNFR1.1 TNFR1	TNFR1.2 TNFR1	TNFR1.1 TNFR1	TNFR1.2 TNFR1	NFKB65.2 NF κ Bp65	IKB I κ B	CREB2 CREB2
E	STR1 Strom-1	STR2.3' Strom-2	STR3 Strom-3	COL1 Coll-1	COL1.3' Coll-1.3'	COL2.1 Coll-2	COL2.2 Coll-2	COL3 Coll-3	COX1 Cox-1	COX2 Cox-2	12LO 12-LO	15LO 15-LO
F	GELA.1 Gel-A	GELB Gel-B	HME Elastase	MTMMP MT-MMP	PUMP1 Matrilysin	TIMP1 TIMP-1	TIMP2 TIMP-2	TIMP3 TIMP-3	ICAM1 ICAM-1	VCAM VCAM	5LO.1 5-LO	CPLA2.2 cPLA2
G	EGF EGF	FGFA FGF acidic	FGFB FGF basic	IGF1 IGF-I	IGFII IGF-II	TGFA TGF α	TGFB TGF β	PDGFB PDGF β	GALNTN Galntonin	GH1 GH-1	GRO GRO1 α	GCR GR
H	MCP1.1 MCP-1	MCP1.1 MCP-1	MIP1A MIP-1 α	MIP1B MIP-1 β	MIF MIF	RANTES RANTES	INOS iNOS	LDLR LDLR	ALU1 IL-10	ALU2 TNFRp70	ALU3 IL-10	POLYA LDLR

A. thaliana controls
 Human controls

Cytokines and related genes
 Transcription factors and related genes
 MMP's and related genes

Chemokines
 Growth factors and related genes
 Other genes

FIG. 1. Ninety-six-element microarray design. The target element name and the corresponding gene are shown in the layout. Some genes have more than one target element to guarantee specificity of signal. For TNF the targets represent decreasing lengths of 1, 0.8, 0.6, 0.4, and 0.2 kb from left to right.

amplified products ranging from 0.2 to 1.2 kb at concentrations of 1 $\mu\text{g}/\mu\text{l}$ or less. No significant difference in the signal levels was observed within this range of target size and only with 0.2-kb length was a signal reduced upon an 8-fold dilution of the 1 $\mu\text{g}/\mu\text{l}$ sample (data not shown). In this study the average length of the targets was 1 kb, with a few exceptions in the range of ≈ 300 bp, arrayed at a concentration of 1 $\mu\text{g}/\mu\text{l}$. Normally one PCR provided sufficient material to fabricate up to 1000 microarray targets.

In considering positional effects in the development of the targets for the microarrays, selection was biased toward the 3' proximal regions, because the signal was reduced if the target fragment was biased toward the 5' end (data not shown). This result was anticipated since the hybridizing probe is prepared by reverse transcription with oligo(dT)-primed mRNA and is richer in 3' proximal sequences. Cross-hybridizations of probes to targets of a gene family were analyzed with the matrix metal-

loproteinases as the example because they can show regions of sequence identities of greater than 70%. With collagenase-1 (Col-1) and collagenase-2 (Col-2) genes as targets with up to 70% sequence identity, and stromelysin-1 (Strom-1) and stromelysin-2 (Strom-2) genes with different degrees of identity, our results showed that a short region of overlap, even with 70–90% sequence identity, produced a low level of cross-hybridization. However, shorter regions of identity spread over the length of the target resulted in cross-hybridization (data not shown). For closely related genes, targets were designed by avoiding long stretches of homology. For members of a gene family two or more target regions were included to discriminate between specificity of signal versus cross-hybridization.

Monitoring Differential Expression in Cultured Cell Lines. In RA tissue, the monocyte/macrophage population plays a prominent role in phagocytic and immunomodulatory activities. Typ-

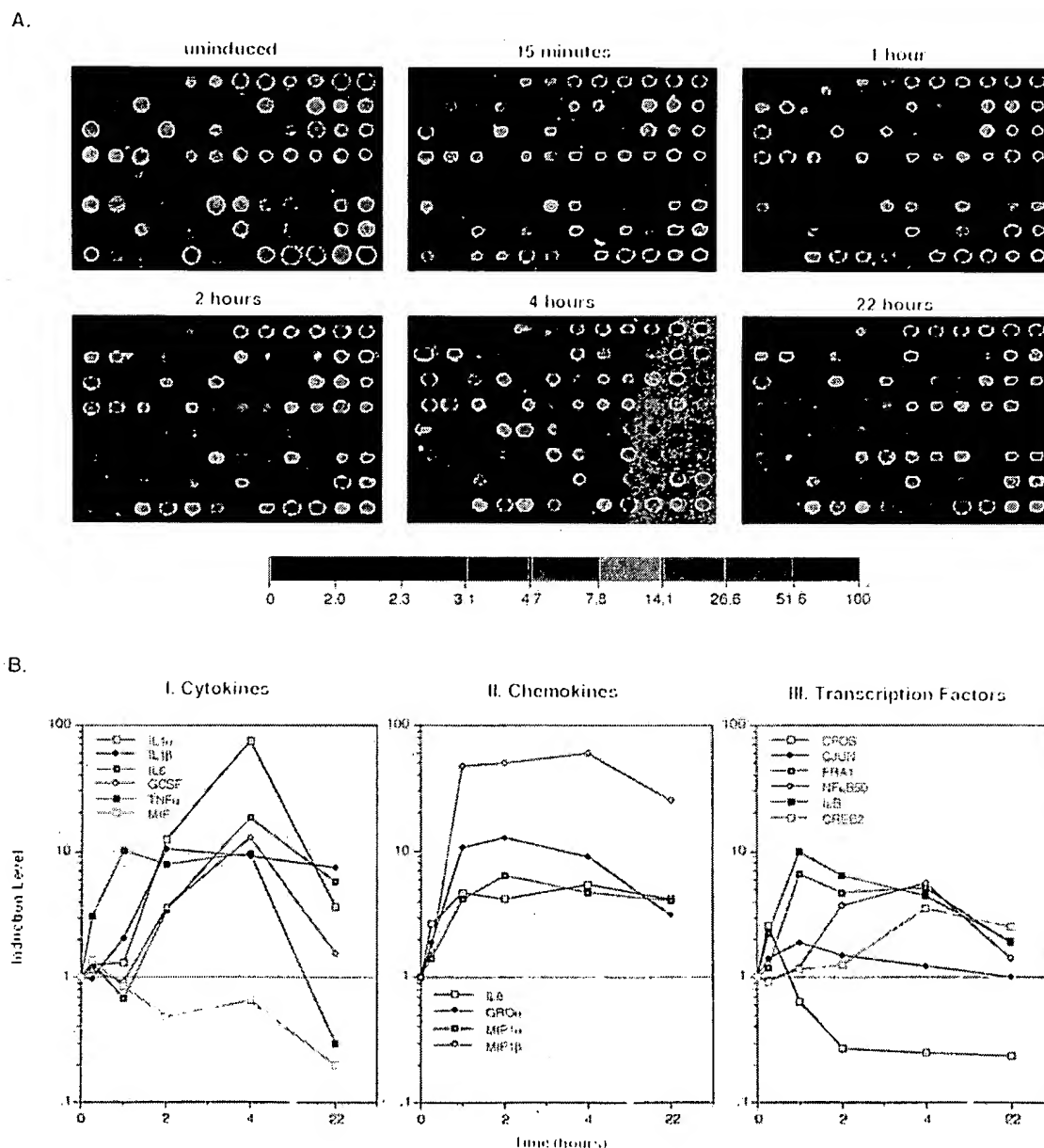


FIG. 2. Time course for LPS/PMA-induced MM6 cells. Array elements are described in Fig. 1. (A) Pseudocolor representations of fluorescent scans correspond to gene expression levels at each time point. The array is made up of 8 *Arabidopsis* control targets and 86 human cDNA targets, the majority of which are genes with known or suspected involvement in inflammation. The color bars provide a comparative calibration scale between arrays and are derived from the *Arabidopsis* mRNA samples that are introduced in equal amounts during probe preparation. Fluorescent probes were made by labeling mRNA from untreated MM6 cells or LPS and PMA treated cells. mRNA was isolated at indicated times after induction. (B I–III) The two-color samples were cohybridized, and microarray scans provided the data for the levels of select transcripts at different time points relative to abundance at time zero. The analysis was performed using normalized data collected from 8-bit images.

ically these cells, when triggered by an immunogen, produce the proinflammatory cytokines TNF and IL-1. We have used the monocyte cell line MM6 and monitored changes in gene expression upon activation with LPS endotoxin, a component of Gram-negative bacterial membranes, and PMA, which augments the action of LPS on TNF production (15). RNA was isolated at different times after induction and used for cDNA probe preparation. From this time course it was clear that TNF expression was induced within 15 min of treatment, reached maximum levels in 1 hr, remained high until 4 hr and subsequently declined (Fig. 2A). Many other cytokine genes were also transiently activated, such as IL-1 α and - β , IL-6, and granulocyte colony-stimulating factor (GCSF). Prominent chemokines activated were IL-8, macrophage inflammatory protein (MIP)-1 β , more so than MIP-1 α , and Gro α or melanoma growth stimulatory factor. Migration inhibitory factor (MIF) expressed in the uninduced state declined in LPS-activated cells. Of the immediate early genes, the noticeable ones were *c-fos*, *fra-1*, *c-jun*, NF- κ Bp50, and I κ B, with *c-rel* expression observed even in the uninduced state (Fig. 2B). These expression patterns are consistent with reported patterns of activation of certain LPS- and PMA-induced genes (12). Demonstrated here is the unique ability of this system to allow parallel visualization of a large number of gene activities over a period of time.

SW1353 cells is a line derived from malignant tumors of the cartilage and behaves much like the chondrocytes upon stimulation with TNF and IL-1 in the expression of MMPs (9). In addition to confirming our earlier observations with Northern blots on Strom-1, Col-1, and Col-3 expression (9), gelatinase (Gel) A, putative metalloproteinase (PUMP)-1 membrane-

type matrix metalloproteinase, tissue inhibitors of matrix metalloproteinases or tissue inhibitor of metalloproteinase 1 (TIMP-1), -2, and -3 were also expressed by these cells together with the human matrix metallo-elastase (HME; Fig. 3A). HME induction was estimated to be \approx 50-fold and was greater than any of the other MMPs examined (Fig. 3B). This result was unexpected because HME is reportedly expressed only by alveolar macrophage and placental cells (16). Expression of the cytokines and chemokines, IL-6, IL-8, MIF, and MIP-1 β was also noted. A variety of other genes, including certain transcription factors, were also up-regulated (Fig. 3), but the overall time-dependent expression of genes in the SW1353 cells was qualitatively distinct from the MM6 cells.

Quantitation of differential gene expression (Figs. 2B and 3B) was achieved with the simultaneous hybridization of Cy3-labeled cDNA from untreated cells and Cy5-labeled cDNA from treated samples. The estimated increases in expression from these microarrays for a select number of genes including IL-1 β , IL-8, MIP-1 β , TNF, HME, Col-1, Col-3, Strom-1, and Strom-2 were compared with data collected from dot blot analysis. Results (not shown) were in close agreement and confirmed our earlier observations on the use of the microarray method for the quantitation of gene expression (3).

Expression Profiles in Primary Chondrocytes and Synoviocytes of Human RA Tissue. Given the sensitivity and the specificity of this method, expression profiles of primary synoviocytes and chondrocytes from diseased tissue were examined. Without prior exposure to inducing agents, low level expression of *c-jun*, GCSF, IL-3, TNF- β , MIF, and RANTES (regulated upon activation, normal T cell expressed and secreted) was seen as well as expression of MMPs, Gela, Strom-1, Col-1, and the three TIMPs. In this case, Col-2 hybridization was considered to be nonspecific because the second Col-2 target taken from the 3' end of the gene gave no

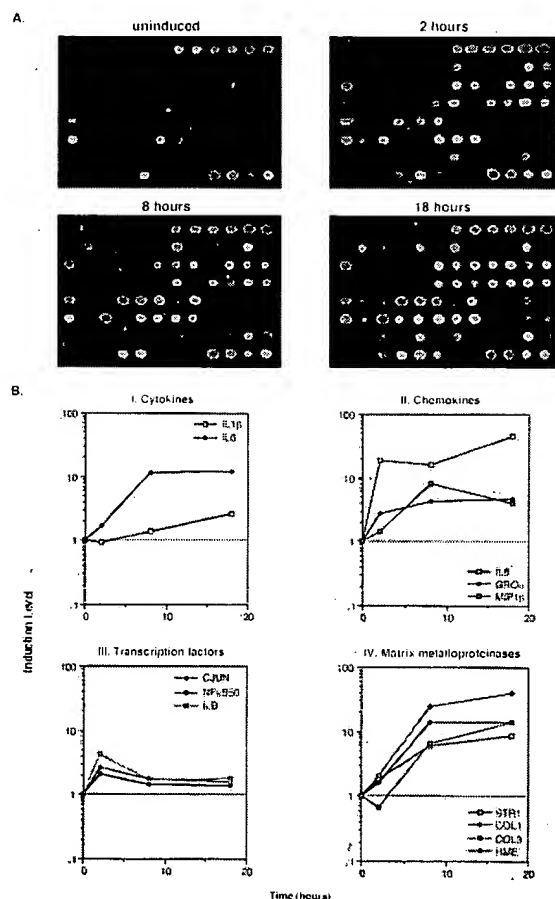


FIG. 3. Time course for IL-1 β and TNF-induced SW1353 cells using the inflammation array (Fig. 1). (A) Pseudocolor representation of fluorescent scans correspond to gene expression levels at each time point. (B I-IV) Relative levels of selected genes at different time points compared with time zero.

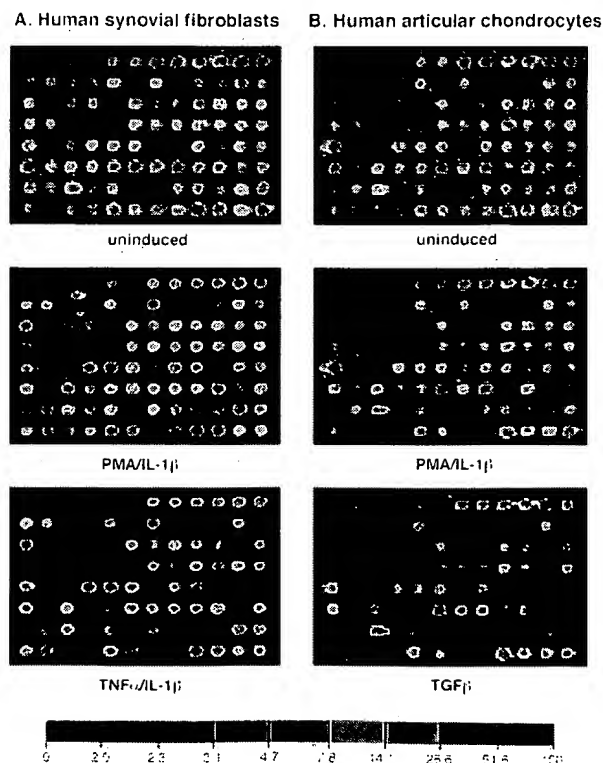


FIG. 4. Expression profiles for early passage primary synoviocytes and chondrocytes isolated from RA tissue, cultured in the presence of 10% fetal calf serum and activated with PMA and IL-1 β , or TNF and IL-1 β , or TGF- β for 18 hr. The color bars provide a comparative calibration scale between arrays and are derived from the *Arabidopsis* mRNA samples that are introduced in equal amounts during probe preparation.

signal. Treatment more so with PMA and IL-1, than TNF and IL-1, produced a dramatic up-regulation in expression of several genes in both of these primary cell types. These genes are as follows: the cytokine IL-6, the chemokines IL-8 and Gro-1 α , and the MMPs; Strom-1, Col-1, Col-3, and HME; and the adhesion molecule, vascular cell adhesion molecule 1 (VCAM-1). The surprise again is HME expression in these primary cells, for reasons discussed above. From these results, the expression profiles of synoviocytes and the chondrocytes appear very similar; the differences are more quantitative than qualitative. Treatment of the primary chondrocytes with the anabolic growth factor TGF- β had an interesting profile in that it produced a remarkable down-regulation of genes expressed in both the untreated and induced state (Fig. 4).

Given the demonstrated effectiveness of this technology, a comparative analysis of two different inflammatory disease states was conducted with probes made from RA tissue and IBD samples. RA samples were from late stage rheumatoid synovial tissue, and IBD specimens were obtained from inflamed lower intestinal mucosa of patients with Crohn disease. With both the 96-element known gene microarray and the 1000-gene microarray of cDNAs selected from a peripheral human blood cell library (3), distinct differences in gene expression patterns were evident. On the 96-gene array, RA tissue samples from different affected individuals gave similar profiles (data not shown) as did different samples from the same individual (Fig. 5). These patterns were notably similar to those observed with primary synoviocytes and chondrocytes (Fig. 4). Included in the list of prominently up-regulated genes are IL-6, the MMPs Strom-1, Col-1, GelA, HME, and in

certain samples PUMP, TIMPs, particularly TIMP-1 and TIMP-3, and the adhesion molecule VCAM. Discernible levels of macrophage chemotactic protein 1 (MCP-1), MIF and RANTES were also noted. IBD samples were in comparison, rather subdued although IL-1 converting enzyme (ICE), TIMP-1, and MIF were notable in all the three different IBD samples examined here. In IBD-A, one of three individual samples, ICE, VCAM, Gro α , and MMP expression was more pronounced than in the others.

We also made use of a peripheral blood cDNA library (3) to identify genes expressed by lymphocytes infiltrating the inflamed tissues from the circulating blood. With the 1046-element array of randomly selected cDNAs from this library, probes made from RA and IBD samples showed hybridizations to a large number of genes. Of these, many were common between the two disease tissues while others were differentially expressed (data not shown). A complete survey of these genes was beyond the scope of this study, but for this report we picked three genes that were up-regulated in the RA tissue relative to IBD. These cDNAs were sequenced and identified by comparison to the GenBank database. They are TIMP-1, apoferritin light chain, and manganese superoxide dismutase (MnSOD). Differential expression of MnSOD was only observed in samples of RA tissue explants maintained in growth medium without serum for anywhere between 2 to 16 hr. These results also indicate that the expression profile of genes can be altered when explants are transferred to culture conditions.

DISCUSSION

The speed, ease, and feasibility of simultaneously monitoring differential expression of hundreds of genes with the cDNA microarray based system (1–3) is demonstrated here in the analysis of a complex disease such as RA. Many different cell types in the RA tissue; macrophages, lymphocytes, plasma cells, neutrophils, synoviocytes, chondrocytes, etc. are known to contribute to the development of the disease with the expression of gene products known to be proinflammatory. They include the cytokines, chemokines, growth factors, MMPs, eicosanoids, and others (7, 11–14), and the design of the 96-element known gene microarray was based on this knowledge and depended on the availability of the genes. The technology was validated by confirming earlier observations on the expression of TNF by the monocyte cell line MM6, and of Col-1 and Col-3 expression in the chondrosarcoma cells and articular chondrocytes (9, 12). In our time-dependent survey the chronological order of gene activities in and between gene families was compared and the results have provided unprecedented profiles of the cytokines (TNF, IL-1, IL-6, GCSF, and MIF), chemokines (MIP-1 α , MIP-1 β , IL-8, and Gro-1), certain transcription factors, and the matrix metalloproteinases (GelA, Strom-1, Col-1, Col-3, HME) in the macrophage cell line MM6 and in the SW1353 chondrosarcoma cells.

Earlier reports of cytokine production in the diseased state had established a model in which TNF is a major participant in RA. Its expression reportedly preceded that of the other cytokines and effector molecules (4). Our results strongly support these results as demonstrated in the time course of the MM6 cells where TNF induction preceded that of IL-1 α and IL-1 β followed by IL-6 and GCSF. These expression profiles demonstrate the utility of the microarrays in determining the hierarchy of signaling events.

In the SW1353 chondrosarcoma cells, all the known MMPs and TIMPs were examined simultaneously. HME expression was discovered, which previously had been observed in only the stromal cells and alveolar macrophages of smoker's lungs and in placental tissue. Its presence in cells of the RA tissue is meaningful because its activity can cause significant destruction of elastin and basement membrane components (16, 17). Expression profiles of synovial fibroblasts and articular chondrocytes were remarkably similar and not too different from the SW1353 cells, indicating that the fibroblast and the chondrocyte can play equally aggressive roles in joint erosion. Prominent genes expressed were

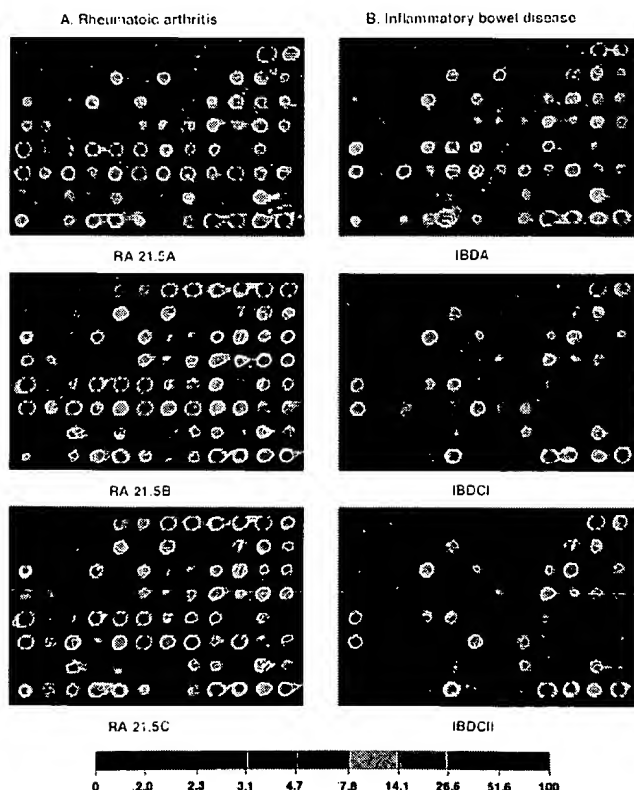


FIG. 5. Expression profiles of RA tissue (A) and IBD tissue (B). mRNA from RA tissue samples obtained from the same individual was isolated directly after excision (RA 21.5A) or maintained in culture without serum for 2 hr (RA 21.5B) or for 6 hr (RA 21.5C). Profiles from tissue samples of two other individuals (data not shown) were remarkably similar to the ones shown here. IBD-A and IBD-CI are from mRNA samples prepared directly after surgery from two separate individuals. For the IBD-CII probe, the tissue sample was cultured in medium without serum for 2 hr before mRNA preparation.

the MMPs, but chemokines and cytokines were also produced by these cells. The effect of the anabolic growth factor TGF- β was profoundly evident in demonstrating the down regulation of these catabolic activities.

RA tissue samples undeniably reflected profiles similar to the cell types examined. Active genes observed were IL-3, IL-6, ICE, the MMPs including HME and TIMPs, chemokines IL-8, Gro α , MIP, MIF, and RANTES, and the adhesion molecule VCAM. Of the growth factors, fibroblast growth factor β was observed most frequently. In comparison, the expression patterns in the other inflammatory state (i.e., IBD) were not as marked as in the RA samples, at least as obtained from the tissue samples selected for this study.

As an alternative approach, the 1046 cDNA microarray of randomly selected genes from a lymphocyte library was used to identify genes expressed in RA tissue (3). Many genes on this array hybridized with probes made from both RA and IBD tissue samples. The results are not surprising because inflammatory tissue is abundantly supplied with cell types infiltrating from the circulating blood, made apparent also by the high levels of chemokine expression in RA tissue. Because of the magnitude of the effort required to identify all the hybridized genes, we have for this report chosen to describe only three differentially expressed genes mainly to verify this method of analysis.

Of the large number of genes observed here, a fair number were already known as active participants in inflammatory disease. These are TNF, IL-1, IL-6, IL-8, GCSF, RANTES, and VCAM. The novel participants not previously reported are HME, IL-3, ICE, and Gro α . With our discovery of HME expression in RA, this gene becomes a target for drug intervention. ICE is a cysteine protease well known for its IL-1 β processing activity (18), and recognized for its role in apoptotic cell death (19). Its expression in RA tissue is intriguing. IL-3 is recognized for its growth-promoting activity in hematopoietic cell lineages, is a product of activated T cells (20), and its expression in synovocytes and chondrocytes of RA tissue is a novel observation.

Like IL-8, Gro α , is a C-X-C subgroup chemokine and is a potent neutrophil and basophil chemoattractant. It down-regulates the expression of types I and III interstitial collagens (21, 22) and is seen here produced by the MM6 cells, in primary synovocytes, and in RA tissue. With the presence of RANTES, MCP, and MIP-1 β , the C-C chemokines (23) migration and infiltration of monocytes, particularly T cells, into the tissue is also enhanced (5) and aid in the trafficking and recruitment of leukocytes into the RA tissue. Their activation, phagocytosis, degranulation, and respiratory bursts could be responsible for the induction of MnSOD in RA. MnSOD is also induced by TNF and IL-1 and serves a protective function against oxidative damage. The induction of the ferritin light chain encoding gene in this tissue may be for reasons similar to those for MnSOD. Ferritin is the major intracellular iron storage protein and it is responsive to intracellular oxidative stress and reactive oxygen intermediates generated during inflammation (24, 25). The active expression of TIMP-1 in RA tissue, as detected by the 1000-element array, is no surprise because our results have repeatedly shown TIMP-1 to be expressed in the constitutive and induced states of RA cells and tissues.

The suitability of the cDNA microarray technology for profiling diseases and for identifying disease related genes is well documented here. This technology could provide new

targets for drug development and disease therapies, and in doing so allow for improved treatment of chronic diseases that are challenging because of their complexity.

We would like to thank the following individuals for their help in obtaining reagents or providing cDNA clones to use as templates in target preparation: N. Arai, P. Cannon, D. R. Cohen, T. Curran, V. Dixit, D. A. Geller, G. I. Goldberg, M. Karin, M. Lotz, L. Matrisian, G. Nolan, C. Lopez-Otin, T. Schall, S. Shapiro, I. Verma, and H. Van Wart. Support for R.W.D., M.S., and R.A.H. was provided by the National Institutes of Health (Grants R37HG00198 and HG00205).

- Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. (1995) *Science* **270**, 467-470.
- Shalon, D., Smith, S. & Brown, P. O. (1996) *Genome Res.* **6**, 639-645.
- Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P. O. & Davis, R. W. (1996) *Proc. Natl. Acad. Sci. USA* **93**, 10614-10619.
- Feldmann, M., Brennan F. M. & Maini, R. N. (1996) *Rheumatoid Arthritis Cell* **85**, 307-310.
- Schall, T. J. (1994) in *The Cytokine Handbook*, ed. Thomson, A. W. (Academic, New York), 2nd Ed., pp. 410-460.
- Lotz, M. F., Blanco, J., Von Kempf, J., Dudler, J., Maier, R., Villiger P. M. & Geng, Y. (1995) *J. Rheumatol.* **22**, Supplement 43, 104-108.
- Birkedal-Hansen, H., Moore, W. G. I., Bodden, M. K., Windsor, L. J., Birkedal-Hansen, B., DeCarlo, A. & Engler, J. A. (1993) *Crit. Rev. Oral Biol. Med.* **4**, 197-250.
- Zeigler-Heitbrock, H. W. L., Thiel, E., Futterer, A., Volker, H., Wirtz, A. & Reithmuller, G. (1988) *Int. J. Cancer* **41**, 456-461.
- Borden, P., Solymar, D., Sucharczuk, A., Lindman, B., Cannon, P. & Heller, R. A. (1996) *J. Biol. Chem.* **271**, 23577-23581.
- Gadher, S. J. & Woolley, D. E. (1987) *Rheumatol. Int.* **7**, 13-22.
- Harris, E. D., Jr. (1990) *New Engl. J. Med.* **322**, 1277-1289.
- Firestein, G. S. (1996) in *Textbook of Rheumatology*, eds. Kelly, W. N., Harris, E. D., Ruddy, S. & Sledge, C. B. (Saunders, Philadelphia), 5th Ed. pp. 5001-5047.
- Alvaro-Garcia, J. M., Zvaifler, Nathan J., Brown, C. B., Kaushansky, K. & Firestein, Gary S. (1991) *J. Immunol.* **146**, 3365-3371.
- Firestein, G. S., Alvaro-Garcia, J. M. & Maki, R. (1990) *J. Immunol.* **144**, 3347-3352.
- Pradines-Figueres, A. & Raetz, C. R. H. (1992) *J. Biol. Chem.* **267**, 23261-23268.
- Shapiro, S. D., Kobayashi, D. L. & Ley, T. J. (1993) *J. Biol. Chem.* **268**, 23824-23829.
- Shipley, M. J., Wesselschmidt, R. L., Kobayashi, D. K., Ley, T. J. & Shapiro, S. D. (1996) *Proc. Natl. Acad. Sci. USA* **93**, 3042-3046.
- Cerreti, D. P., Kozlosky, C. J., Mosley, B., Nelson, N., Van Ness, K., Greenstreet, T. A., March, C. J., Kronheim, S. R., Druck, T., Cannizaro, L. A., Huebner, K. & Black, R. A. (1992) *Science* **256**, 97-100.
- Miura, M., Zhu, H., Rotello, R., Hartweig, E. A. & Yuan, J. (1993) *Cell* **75**, 653-660.
- Arai, K., Lee, F., Miyajima, A., Shoichiro, M., Arai, N. & Takashi, Y. (1990) *Annu. Rev. Biochem.* **59**, 783-836.
- Geiser, T., Dewald, B., Ehrenguber, M. U., Lewis, I. C. & Baggiolini, M. (1993) *J. Biol. Chem.* **268**, 15419-15424.
- Unemori, E. N., Amento, E. P., Bauer, E. A. & Horuk, R. (1993) *J. Biol. Chem.* **268**, 1338-1342.
- Robinson, E., Keystone, E. C., Schall, T. J., Gillet, N. & Fish, E. N. (1995) *Clin. Exp. Immunol.* **101**, 398-407.
- Roeser, H. (1980) in *Iron Metabolism in Biochemistry and Medicine*, eds. Jacobs, A. & Worwood, M. (Academic, New York), Vol. 2, pp. 605-640.
- Kwak, E. L., Larochelle, D. A., Beaumont, C., Torti, S. V. & Torti, F. M. (1995) *J. Biol. Chem.* **270**, 15285-15293.



PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification ⁶ : C12Q 1/68, C07H 21/04	A1	(11) International Publication Number: WO 97/13877 (43) International Publication Date: 17 April 1997 (17.04.97)
(21) International Application Number: PCT/US96/16342 (22) International Filing Date: 11 October 1996 (11.10.96) (30) Priority Data: PCT/US95/12791 12 October 1995 (12.10.95) WO (34) Countries for which the regional or international application was filed: US et al. PCT/US96/09513 6 June 1996 (06.06.96) WO (34) Countries for which the regional or international application was filed: US et al. (60) Parent Application or Grant (63) Related by Continuation US Not furnished (CIP) Filed on Not furnished (71) Applicant (for all designated States except US): LYNX THERAPEUTICS, INC. [US/US]; 3832 Bay Center Place, Hayward, CA 94545 (US). (72) Inventor; and (75) Inventor/Applicant (for US only): MARTIN, David, W. [US/US]; Lynx Therapeutics, Inc., 3832 Bay Center Place, Hayward, CA 94545 (US).		(74) Agent: POWERS, Vincent, M.; Dehlinger & Associates, Post Office Box 60850, Palo Alto, CA 94306-0850 (US). (81) Designated States: AU, CA, CZ, EE, FI, HU, JP, KR, LT, LV, NO, NZ, PL, RU, SG, US, European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE). Published With international search report. Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.
(54) Title: MEASUREMENT OF GENE EXPRESSION PROFILES IN TOXICITY DETERMINATION (57) Abstract <p>A method is provided for assessing the toxicity of a compound in a test organism by measuring gene expression profiles of selected tissues. Gene expression profiles are measured by massively parallel signature sequencing of cDNA libraries constructed from mRNA extracted from the selected tissues. Gene expression profiles provide extensive information on the effects of administering a compound to a test organism in both acute toxicity tests and in prolonged and chronic toxicity tests.</p>		

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AM	Armenia	GB	United Kingdom	MW	Malawi
AT	Austria	GE	Georgia	MX	Mexico
AU	Australia	GN	Guinea	NE	Niger
BB	Barbados	GR	Greece	NL	Netherlands
BE	Belgium	HU	Hungary	NO	Norway
BF	Burkina Faso	IE	Ireland	NZ	New Zealand
BG	Bulgaria	IT	Italy	PL	Poland
BJ	Benin	JP	Japan	PT	Portugal
BR	Brazil	KE	Kenya	RO	Romania
BY	Belarus	KG	Kyrgyzstan	RU	Russian Federation
CA	Canada	KP	Democratic People's Republic of Korea	SD	Sudan
CF	Central African Republic	KR	Republic of Korea	SE	Sweden
CG	Congo	KZ	Kazakhstan	SG	Singapore
CH	Switzerland	LI	Liechtenstein	SI	Slovenia
CI	Côte d'Ivoire	LK	Sri Lanka	SK	Slovakia
CM	Cameroon	LR	Liberia	SN	Senegal
CN	China	LT	Lithuania	SZ	Swaziland
CS	Czechoslovakia	LU	Luxembourg	TD	Chad
CZ	Czech Republic	LV	Latvia	TG	Togo
DE	Germany	MC	Monaco	TJ	Tajikistan
DK	Denmark	MD	Republic of Moldova	TT	Trinidad and Tobago
EE	Estonia	MG	Madagascar	UA	Ukraine
ES	Spain	ML	Mali	UG	Uganda
FI	Finland	MN	Mongolia	US	United States of America
FR	France	MR	Mauritania	UZ	Uzbekistan
GA	Gabon			VN	Viet Nam

MEASUREMENT OF GENE EXPRESSION PROFILES
IN TOXICITY DETERMINATION

5

Field of the Invention

The invention relates generally to methods for detecting and monitoring phenotypic changes in in vitro and in vivo systems for assessing and/or determining the toxicity of chemical compounds, and more particularly, the invention relates to a method for detecting and monitoring changes in gene expression patterns in in vitro and in vivo systems for determining the toxicity of drug candidates.

BACKGROUND

The ability to rapidly and conveniently assess the toxicity of new compounds is extremely important. Thousands of new compounds are synthesized every year, and many are introduced to the environment through the development of new commercial products and processes, often with little knowledge of their short term and long term health effects. In the development of new drugs, the cost of assessing the safety and efficacy of candidate compounds is becoming astronomical: It is estimated that the pharmaceutical industry spends an average of about 300 million dollars to bring a new pharmaceutical compound to market, e.g. Biotechnology, 13: 226-228 (1995). A large fraction of these costs are due to the failure of candidate compounds in the later stages of the developmental process. That is, as the assessment of a candidate drug progresses from the identification of a compound as a drug candidate--for example, through relatively inexpensive binding assays or in vitro screening assays, to pharmacokinetic studies, to toxicity studies, to efficacy studies in model systems, to preliminary clinical studies, and so on, the costs of the associated tests and analyses increases tremendously. Consequently, it may cost several tens of millions of dollars to determine that a once promising candidate compound possesses a side effect or cross reactivity that renders it commercially infeasible to develop further. A great challenge of pharmaceutical development is to remove from further consideration as early as possible those compounds that are likely to fail in the later stages of drug testing.

Drug development programs are clearly structured with this objective in mind; however, rapidly escalating costs have created a need to develop even more stringent and less expensive screens in the early stages to identify false leads as soon as possible. Toxicity assessment is an area where such improvements may be made, for both drug development and for assessing the environmental, health, and safety effects of new compounds in general.

Typically the toxicity of a compound is determined by administering the compound to one or more species of test animal under controlled conditions and by monitoring the effects on a wide range of parameters. The parameters include such things as blood chemistry, weight gain or loss, a variety of behavioral patterns, muscle tone, body temperature, respiration rate, lethality, and the like, which collectively provide a measure of the state of health of the test animal. The degree of deviation of such parameters from their normal ranges gives a measure of the toxicity of a compound. Such tests may be designed to assess the acute, prolonged, or chronic toxicity of a compound. In general, acute tests involve administration of the test chemical on one occasion. The period of observation of the test animals may be as short as a few hours, although it is usually at least 24 hours and in some cases it may be as long as a week or more. In general, prolonged tests involve administration of the test chemical on multiple occasions. The test chemical may be administered one or more times each day, irregularly as when it is incorporated in the diet, at specific times such as during pregnancy, or in some cases regularly but only at weekly intervals. Also, in the prolonged test the experiment is usually conducted for not less than 90 days in the rat or mouse or a year in the dog. In contrast to the acute and prolonged types of test, the chronic toxicity tests are those in which the test chemical is administered for a substantial portion of the lifetime of the test animal. In the case of the mouse or rat, this is a period of 2 to 3 years. In the case of the dog, it is for 5 to 7 years.

Significant costs are incurred in establishing and maintaining large cohorts of test animals for such assays, especially the larger animals in chronic toxicity assays. Moreover, because of species specific effects, passing such toxicity tests does not ensure that a compound is free of toxic effects when used in humans. Such tests do, however, provide a standardized set of information for judging the safety of new compounds, and they provide a database for giving preliminary assessments of related compounds. An important area for improving toxicity determination would be the identification of new observables which are predictive of the outcome of the expensive and tedious animal assays.

In other medical fields, there has been significant interest in applying recent advances in biotechnology, particularly in DNA sequencing, to the identification and study of differentially expressed genes in healthy and diseased organisms, e.g. Adams et al, Science, 252: 1651-1656 (1991); Matsubara et al, Gene, 135: 265-274 (1993); Rosenberg et al, International patent application, PCT/US95/01863. The objectives of such applications include increasing our knowledge of disease processes, identifying genes that play important roles in the disease process, and providing diagnostic and therapeutic approaches that exploit the expressed genes or their

products. While such approaches are attractive, those based on exhaustive, or even sampled, sequencing of expressed genes are still beset by the enormous effort required: It is estimated that 30-35 thousand different genes are expressed in a typical mammalian tissue in any given state, e.g. Ausubel et al, Editors, Current Protocols, 5.8.1-5.8.4 (John Wiley & Sons, New York, 1992). Determining the sequences of even a small sample of that number of gene products is a major enterprise, requiring industrial-scale resources. Thus, the routine application of massive sequencing of expressed genes is still beyond current commercial technology.

The availability of new assays for assessing the toxicity of compounds, such as candidate drugs, that would provide more comprehensive and precise information about the state of health of a test animal would be highly desirable. Such additional assays would preferably be less expensive, more rapid, and more convenient than current testing procedures, and would at the same time provide enough information to make early judgments regarding the safety of new compounds.

15

Summary of the Invention

An object of the invention is to provide a new approach to toxicity assessment based on an examination of gene expression patterns, or profiles, in in vitro or in vivo test systems.

Another object of the invention is to provide a database on which to base decisions concerning the toxicological properties of chemicals, particularly drug candidates.

A further object of the invention is to provide a method for analyzing gene expression patterns in selected tissues of test animals.

A still further object of the invention is to provide a system for identifying genes which are differentially expressed in response to exposure to a test compound.

Another object of the invention is to provide a rapid and reliable method for correlating gene expression with short term and long term toxicity in test animals.

Another object of the invention is to identify genes whose expression is predictive of deleterious toxicity.

The invention achieves these and other objects by providing a method for massively parallel signature sequencing of genes expressed in one or more selected tissues of an organism exposed to a test compound. An important feature of the invention is the application of novel DNA sorting and sequencing methodologies that permit the formation of gene expression profiles for selected tissues by determining the sequence of portions of many thousands of different polynucleotides in parallel. Such profiles may be compared with those from tissues of control organisms at single or multiple time points to identify expression patterns predictive of toxicity.

The sorting methodology of the invention makes use of oligonucleotide tags that are members of a minimally cross-hybridizing set of oligonucleotides. The sequences of oligonucleotides of such a set differ from the sequences of every other member of the same set by at least two nucleotides. Thus, each member of such a set cannot form a duplex (or triplex) with the complement of any other member with less than two mismatches. Complements of oligonucleotide tags of the invention, referred to herein as "tag complements," may comprise natural nucleotides or non-natural nucleotide analogs. Preferably, tag complements are attached to solid phase supports. Such oligonucleotide tags when used with their corresponding tag complements provide a means of enhancing specificity of hybridization for sorting polynucleotides, such as cDNAs.

The polynucleotides to be sorted each have an oligonucleotide tag attached, such that different polynucleotides have different tags. As explained more fully below, this condition is achieved by employing a repertoire of tags substantially greater than the population of polynucleotides and by taking a sufficiently small sample of tagged polynucleotides from the full ensemble of tagged polynucleotides. After such sampling, when the populations of supports and polynucleotides are mixed under conditions which permit specific hybridization of the oligonucleotide tags with their respective complements, identical polynucleotides sort onto particular beads or regions. The sorted populations of polynucleotides can then be sequenced on the solid phase support by a "single-base" or "base-by-base" sequencing methodology, as described more fully below.

In one aspect, the method of the invention comprises the following steps: (a) administering the compound to a test organism; (b) extracting a population of mRNA molecules from each of one or more tissues of the test organism; (c) forming a separate population of cDNA molecules from each population of mRNA molecules extracted from the one or more tissues such that each cDNA molecule of the separate populations has an oligonucleotide tag attached, the oligonucleotide tags being selected from the same minimally cross-hybridizing set; (d) separately sampling each population of cDNA molecules such that substantially all different cDNA molecules within a separate population have different oligonucleotide tags attached; (e) sorting the cDNA molecules of each separate population by specifically hybridizing the oligonucleotide tags with their respective complements, the respective complements being attached as uniform populations of substantially identical complements in spatially discrete regions on one or more solid phase supports; (f) determining the nucleotide sequence of a portion of each of the sorted cDNA molecules of each separate population to form a frequency distribution of expressed genes for each of

the one or more tissues; and (g) correlating the frequency distribution of expressed genes in each of the one or more tissues with the toxicity of the compound.

An important aspect of the invention is the identification of genes whose expression is predictive of the toxicity of a compound. Once such genes are identified, they may be employed in conventional assays, such as reverse transcriptase polymerase chain reaction (RT-PCR) assays for gene expression.

Brief Description of the Drawings

Figure 1 is a flow chart representation of an algorithm for generating minimally cross-hybridizing sets of oligonucleotides.

Figure 2 diagrammatically illustrates an apparatus for carrying out polynucleotide sequencing in accordance with the invention.

Definitions

"Complement" or "tag complement" as used herein in reference to oligonucleotide tags refers to an oligonucleotide to which a oligonucleotide tag specifically hybridizes to form a perfectly matched duplex or triplex. In embodiments where specific hybridization results in a triplex, the oligonucleotide tag may be selected to be either double stranded or single stranded. Thus, where triplexes are formed, the term "complement" is meant to encompass either a double stranded complement of a single stranded oligonucleotide tag or a single stranded complement of a double stranded oligonucleotide tag.

The term "oligonucleotide" as used herein includes linear oligomers of natural or modified monomers or linkages, including deoxyribonucleosides, ribonucleosides, anomeric forms thereof, peptide nucleic acids (PNAs), and the like, capable of specifically binding to a target polynucleotide by way of a regular pattern of monomer-to-monomer interactions, such as Watson-Crick type of base pairing, base stacking, Hoogsteen or reverse Hoogsteen types of base pairing, or the like. Usually monomers are linked by phosphodiester bonds or analogs thereof to form oligonucleotides ranging in size from a few monomeric units, e.g. 3-4, to several tens of monomeric units. Whenever an oligonucleotide is represented by a sequence of letters, such as "ATGCCTG," it will be understood that the nucleotides are in 5'→3' order from left to right and that "A" denotes deoxyadenosine, "C" denotes deoxycytidine, "G" denotes deoxyguanosine, and "T" denotes thymidine, unless otherwise noted. Analogs of phosphodiester linkages include phosphorothioate, phosphorodithioate, phosphoranilidate, phosphoramidate, and the like. Usually oligonucleotides of the invention comprise the four natural nucleotides; however, they may also comprise non-natural nucleotide analogs. It is clear to those skilled in the

art when oligonucleotides having natural or non-natural nucleotides may be employed, e.g. where processing by enzymes is called for, usually oligonucleotides consisting of natural nucleotides are required.

"Perfectly matched" in reference to a duplex means that the poly- or oligonucleotide strands making up the duplex form a double stranded structure with one other such that every nucleotide in each strand undergoes Watson-Crick basepairing with a nucleotide in the other strand. The term also comprehends the pairing of nucleoside analogs, such as deoxyinosine, nucleosides with 2-aminopurine bases, and the like, that may be employed. In reference to a triplex, the term means that the triplex consists of a perfectly matched duplex and a third strand in which every nucleotide undergoes Hoogsteen or reverse Hoogsteen association with a basepair of the perfectly matched duplex. Conversely, a "mismatch" in a duplex between a tag and an oligonucleotide means that a pair or triplet of nucleotides in the duplex or triplex fails to undergo Watson-Crick and/or Hoogsteen and/or reverse Hoogsteen bonding.

As used herein, "nucleoside" includes the natural nucleosides, including 2'-deoxy and 2'-hydroxyl forms, e.g. as described in Kornberg and Baker, DNA Replication, 2nd Ed. (Freeman, San Francisco, 1992). "Analog" in reference to nucleosides includes synthetic nucleosides having modified base moieties and/or modified sugar moieties, e.g. described by Scheit, Nucleotide Analogs (John Wiley, New York, 1980); Uhlman and Peyman, Chemical Reviews, 90: 543-584 (1990), or the like, with the only proviso that they are capable of specific hybridization. Such analogs include synthetic nucleosides designed to enhance binding properties, reduce complexity, increase specificity, and the like.

As used herein "sequence determination" or "determining a nucleotide sequence" in reference to polynucleotides includes determination of partial as well as full sequence information of the polynucleotide. That is, the term includes sequence comparisons, fingerprinting, and like levels of information about a target polynucleotide, as well as the express identification and ordering of nucleosides, usually each nucleoside, in a target polynucleotide. The term also includes the determination of the identification, ordering, and locations of one, two, or three of the four types of nucleotides within a target polynucleotide. For example, in some embodiments sequence determination may be effected by identifying the ordering and locations of a single type of nucleotide, e.g. cytosines, within the target polynucleotide "CATCGC ..." so that its sequence is represented as a binary code, e.g. "100101 ..." for "C-(not C)-(not C)-C-(not C)-C ..." and the like.

As used herein, the term "complexity" in reference to a population of polynucleotides means the number of different species of molecule present in the population.

As used herein, the terms "gene expression profile," and "gene expression pattern" which is used equivalently, means a frequency distribution of sequences of portions of cDNA molecules sampled from a population of tag-cDNA conjugates. Generally, the portions of sequence are sufficiently long to uniquely identify the cDNA from which the portion arose. Preferably, the total number of sequences determined is at least 1000; more preferably, the total number of sequences determined in a gene expression profile is at least ten thousand.

As used herein, "test organism" means any in vitro or in vivo system which provides measureable responses to exposure to test compounds. Typically, test organisms may be mammalian cell cultures, particularly of specific tissues, such as hepatocytes, neurons, kidney cells, colony forming cells, or the like, or test organisms may be whole animals, such as rats, mice, hamsters, guinea pigs, dogs, cats, rabbits, pigs, monkeys, and the like.

Detailed Description of the Invention

The invention provides a method for determining the toxicity of a compound by analyzing changes in the gene expression profiles in selected tissues of test organisms exposed to the compound. The invention also provides a method of identifying toxicity markers consisting of individual genes or a group of genes that is expressed acutely and which is correlated with prolonged or chronic toxicity, or suggests that the compound will have an undesirable cross reactivity. Gene expression profiles are generated by sequencing portions of cDNA molecules construction from mRNA extracted from tissues of test organisms exposed to the compound being tested. As used herein, the term "tissue" is employed with its usual medical or biological meaning, except that in reference to an in vitro test system, such as a cell culture, it simply means a sample from the culture. Gene expression profiles derived from test organisms are compared to gene expression profiles derived from control organisms to determine the genes which are differentially expressed in the test organism because of exposure to the compound being tested. In both cases, the sequence information of the gene expression profiles is obtained by massively parallel signature sequencing of cDNAs, which is implemented in steps (c) through (f) of the above method.

Toxicity Assessment

Procedures for designing and conducting toxicity tests in in vitro and in vivo systems is well known, and is described in many texts on the subject, such as Loomis

et al. Loomis's Essentials of Toxicology, 4th Ed. (Academic Press, New York, 1996); Echobichon, The Basics of Toxicity Testing (CRC Press, Boca Raton, 1992); Frazier, editor, In Vitro Toxicity Testing (Marcel Dekker, New York, 1992); and the like.

In toxicity testing, two groups of test organisms are usually employed: one group serves as a control and the other group receives the test compound in a single dose (for acute toxicity tests) or a regimen of doses (for prolonged or chronic toxicity tests). Since in most cases, the extraction of tissue as called for in the method of the invention requires sacrificing the test animal, both the control group and the group receiving compound must be large enough to permit removal of animals for sampling tissues, if it is desired to observe the dynamics of gene expression through the duration of an experiment.

In setting up a toxicity study, extensive guidance is provided in the literature for selecting the appropriate test organism for the compound being tested, route of administration, dose ranges, and the like. Water or physiological saline (0.9% NaCl in water) is the solute of choice for the test compound since these solvents permit administration by a variety of routes. When this is not possible because of solubility limitations, it is necessary to resort to the use of vegetable oils such as corn oil or even organic solvents, of which propylene glycol is commonly used. Whenever possible the use of suspension or emulsion should be avoided except for oral administration. Regardless of the route of administration, the volume required to administer a given dose is limited by the size of the animal that is used. It is desirable to keep the volume of each dose uniform within and between groups of animals. When rats or mice are used the volume administered by the oral route should not exceed 0.005 ml per gram of animal. Even when aqueous or physiological saline solutions are used for parenteral injection the volumes that are tolerated are limited, although such solutions are ordinarily thought of as being innocuous. The intravenous LD₅₀ of distilled water in the mouse is approximately 0.044 ml per gram and that of isotonic saline is 0.068 ml per gram of mouse.

When a compound is to be administered by inhalation, special techniques for generating test atmospheres are necessary. Dose estimation becomes very complicated. The methods usually involve aerosolization or nebulization of fluids containing the compound. If the agent to be tested is a fluid that has an appreciable vapor pressure, it may be administered by passing air through the solution under controlled temperature conditions. Under these conditions, dose is estimated from the volume of air inhaled per unit time, the temperature of the solution, and the vapor pressure of the agent involved. Gases are metered from reservoirs. When particles of a solution are to be administered, unless the particle size is less than about 2 μ m the particles will not reach the terminal alveolar sacs in the lungs. A variety of

apparatuses and chambers are available to perform studies for detecting effects of irritant or other toxic endpoints when they are administered by inhalation. The preferred method of administering an agent to animals is via the oral route, either by intubation or by incorporating the agent in the feed.

5 Preferably, in designing a toxicity assessment, two or more species should be employed that handle the test compound as similarly to man as possible in terms of metabolism, absorption, excretion, tissue storage, and the like. Preferably, multiple doses or regimens at different concentrations should be employed to establish a dose-response relationship with respect to toxic effects. And preferably, the route of
10 administration to the test animal should be the same as, or as similar as possible to, the route of administration of the compound to man. Effects obtained by one route of administration to test animals are not a priori applicable to effects by another route of administration to man. For example, food additives for man should be tested by admixture of the material in the diet of the test animals.

15 Acute toxicity tests consist of administering a compound to test organisms on one occasion. The purpose of such test is to determine the symptomatology consequent to administration of the compound and to determine the degree of lethality of the compound. The initial procedure is to perform a series of range-finding doses of the compound in a single species. This necessitates selection of a route of
20 administration, preparation of the compound in a form suitable for administration by the selected route, and selection of an appropriate species. Preferably, initial acute toxicity studies are performed on either rats or mice because of their low cost, their availability, and the availability of abundant toxicologic reference data on these species. Prolonged toxicity tests consist of administering a compound to test
25 organisms repeatedly, usually on a daily basis, over a period of 3 to 4 months. Two practical factors are encountered that place constraints on the design of such tests: First, the available routes of administration are limited because the route selected must be suitable for repeated administration without inducing harmful effects. And second, blood, urine, and perhaps other samples, should be taken repeatedly without
30 inducing significant harm to the test animals. Preferably, in the method of the invention the gene expression profiles are obtained in conjunction with the measurement of the traditional toxicologic parameters, such as listed in the table below:

35

Hematology	Blood Chemistry	Urine Analyses
erythrocyte count	sodium	pH
total leukocyte count	potassium	specific gravity
differential leukocyte count	chloride	total protein
hematocrit	calcium	sediment
hemoglobin	carbon dioxide	glucose
	serum glutamine-pyruvate transaminase	ketones
	serum glutamin-oxalacetic transaminase	bilirubin
	serum protein	
	electrophoresis	
	blood sugar	
	blood urea nitrogen	
	total serum protein	
	serum albumin	
	total serum bilirubin	

Oligonucleotide Tags and Tag Complements

5 Oligonucleotide Tags and Tag Complements

Oligonucleotide tags are members of a minimally cross-hybridizing set of oligonucleotides. The sequences of oligonucleotides of such a set differ from the sequences of every other member of the same set by at least two nucleotides. Thus, each member of such a set cannot form a duplex (or triplex) with the complement of
10 any other member with less than two mismatches. Complements of oligonucleotide tags, referred to herein as "tag complements," may comprise natural nucleotides or non-natural nucleotide analogs. Preferably, tag complements are attached to solid phase supports. Such oligonucleotide tags when used with their corresponding tag complements provide a means of enhancing specificity of hybridization for sorting,
15 tracking, or labeling molecules, especially polynucleotides.

Minimally cross-hybridizing sets of oligonucleotide tags and tag complements may be synthesized either combinatorially or individually depending on the size of the set desired and the degree to which cross-hybridization is sought to be minimized (or stated another way, the degree to which specificity is sought to be enhanced). For example, a minimally cross-hybridizing set may consist of a set of individually synthesized 10-mer sequences that differ from each other by at least 4 nucleotides. such set having a maximum size of 332 (when composed of 3 kinds of nucleotides and counted using a computer program such as disclosed in Appendix 1c). Alternatively, a minimally cross-hybridizing set of oligonucleotide tags may also be

assembled combinatorially from subunits which themselves are selected from a minimally cross-hybridizing set. For example, a set of minimally cross-hybridizing 12-mers differing from one another by at least three nucleotides may be synthesized by assembling 3 subunits selected from a set of minimally cross-hybridizing 4-mers that each differ from one another by three nucleotides. Such an embodiment gives a maximally sized set of 9^3 , or 729, 12-mers. The number 9 is number of oligonucleotides listed by the computer program of Appendix Ia, which assumes, as with the 10-mers, that only 3 of the 4 different types of nucleotides are used. The set is described as "maximal" because the computer programs of Appendices Ia-c provide the largest set for a given input (e.g. length, composition, difference in number of nucleotides between members). Additional minimally cross-hybridizing sets may be formed from subsets of such calculated sets.

Oligonucleotide tags may be single stranded and be designed for specific hybridization to single stranded tag complements by duplex formation or for specific hybridization to double stranded tag complements by triplex formation. Oligonucleotide tags may also be double stranded and be designed for specific hybridization to single stranded tag complements by triplex formation.

When synthesized combinatorially, an oligonucleotide tag preferably consists of a plurality of subunits, each subunit consisting of an oligonucleotide of 3 to 9 nucleotides in length wherein each subunit is selected from the same minimally cross-hybridizing set. In such embodiments, the number of oligonucleotide tags available depends on the number of subunits per tag and on the length of the subunits. The number is generally much less than the number of all possible sequences the length of the tag, which for a tag n nucleotides long would be 4^n .

Complements of oligonucleotide tags attached to a solid phase support are used to sort polynucleotides from a mixture of polynucleotides each containing a tag. Complements of the oligonucleotide tags are synthesized on the surface of a solid phase support, such as a microscopic bead or a specific location on an array of synthesis locations on a single support, such that populations of identical sequences are produced in specific regions. That is, the surface of each support, in the case of a bead, or of each region, in the case of an array, is derivatized by only one type of complement which has a particular sequence. The population of such beads or regions contains a repertoire of complements with distinct sequences. As used herein in reference to oligonucleotide tags and tag complements, the term "repertoire" means the set of minimally cross-hybridizing set of oligonucleotides that make up the tags in a particular embodiment or the corresponding set of tag complements.

The polynucleotides to be sorted each have an oligonucleotide tag attached, such that different polynucleotides have different tags. As explained more fully

below, this condition is achieved by employing a repertoire of tags substantially greater than the population of polynucleotides and by taking a sufficiently small sample of tagged polynucleotides from the full ensemble of tagged polynucleotides. After such sampling, when the populations of supports and polynucleotides are mixed
 5 under conditions which permit specific hybridization of the oligonucleotide tags with their respective complements, identical polynucleotides sort onto particular beads or regions.

The nucleotide sequences of oligonucleotides of a minimally cross-hybridizing set are conveniently enumerated by simple computer programs, such as those
 10 exemplified by programs whose source codes are listed in Appendices Ia and Ib. Program minhx of Appendix Ia computes all minimally cross-hybridizing sets having 4-mer subunits composed of three kinds of nucleotides. Program tagN of Appendix Ib enumerates longer oligonucleotides of a minimally cross-hybridizing set. Similar algorithms and computer programs are readily written for listing oligonucleotides of
 15 minimally cross-hybridizing sets for any embodiment of the invention. Table I below provides guidance as to the size of sets of minimally cross-hybridizing oligonucleotides for the indicated lengths and number of nucleotide differences. The above computer programs were used to generate the numbers.

20

Table I

Oligonucleotide Word Length	Nucleotide Difference between Oligonucleotides of Minimally Cross- Hybridizing Set	Maximal Size of Minimally Cross- Hybridizing Set	Size of Repertoire with Four Words	Size of Repertoire with Five Words
4	3	9	6561	5.90×10^4
6	3	27	5.3×10^5	1.43×10^7
7	4	27	5.3×10^5	1.43×10^7
7	5	8	4096	3.28×10^4
8	3	190	1.30×10^9	2.48×10^{11}
8	4	62	1.48×10^7	9.16×10^8
8	5	18	1.05×10^5	1.89×10^6
9	5	39	2.31×10^6	9.02×10^7
10	5	332	1.21×10^{10}	
10	6	28	6.15×10^5	1.72×10^7
11	5	187		
18	6	≈ 25000		

18

12

24

For some embodiments of the invention, where extremely large repertoires of tags are not required, oligonucleotide tags of a minimally cross-hybridizing set may be separately synthesized. Sets containing several hundred to several thousands, or
5 even several tens of thousands, of oligonucleotides may be synthesized directly by a variety of parallel synthesis approaches, e.g. as disclosed in Frank et al, U.S. patent 4,689,405; Frank et al, Nucleic Acids Research, 11: 4365-4377 (1983); Matson et al, Anal. Biochem., 224: 110-116 (1995); Fodor et al, International application PCT/US93/04145; Pease et al, Proc. Natl. Acad. Sci., 91: 5022-5026 (1994);
10 Southern et al, J. Biotechnology, 35: 217-227 (1994), Brennan, International application PCT/US94/05896; Lashkari et al, Proc. Natl. Acad. Sci., 92: 7912-7915 (1995); or the like.

Preferably, oligonucleotide tags of the invention are synthesized combinatorially out of subunits between three and six nucleotides in length and
15 selected from the same minimally cross-hybridizing set. For oligonucleotides in this range, the members of such sets may be enumerated by computer programs based on the algorithm of Fig. 1.

The algorithm of Fig. 1 is implemented by first defining the characteristics of the subunits of the minimally cross-hybridizing set, i.e. length, number of base
20 differences between members, and composition, e.g. do they consist of two, three, or four kinds of bases. A table M_n , $n=1$, is generated (100) that consists of all possible sequences of a given length and composition. An initial subunit S_1 is selected and compared (120) with successive subunits S_i for $i=n+1$ to the end of the table. Whenever a successive subunit has the required number of mismatches to be a
25 member of the minimally cross-hybridizing set, it is saved in a new table M_{n+1} (125), that also contains subunits previously selected in prior passes through step 120. For example, in the first set of comparisons, M_2 will contain S_1 ; in the second set of comparisons, M_3 will contain S_1 and S_2 ; in the third set of comparisons, M_4 will contain S_1 , S_2 , and S_3 ; and so on. Similarly, comparisons in table M_j will be
30 between S_j and all successive subunits in M_j . Note that each successive table M_{n+1} is smaller than its predecessors as subunits are eliminated in successive passes through step 130. After every subunit of table M_n has been compared (140) the old table is replaced by the new table M_{n+1} , and the next round of comparisons are begun. The process stops (160) when a table M_n is reached that contains no
35 successive subunits to compare to the selected subunit S_i , i.e. $M_n=M_{n+1}$.

Preferably, minimally cross-hybridizing sets comprise subunits that make approximately equivalent contributions to duplex stability as every other subunit in

the set. In this way, the stability of perfectly matched duplexes between every subunit and its complement is approximately equal. Guidance for selecting such sets is provided by published techniques for selecting optimal PCR primers and calculating duplex stabilities, e.g. Rychlik et al, Nucleic Acids Research, 17: 8543-8551 (1989) and 18: 6409-6412 (1990); Breslauer et al, Proc. Natl. Acad. Sci., 83: 3746-3750 (1986); Wetmur, Crit. Rev. Biochem. Mol. Biol., 26: 227-259 (1991); and the like. For shorter tags, e.g. about 30 nucleotides or less, the algorithm described by Rychlik and Wetmur is preferred, and for longer tags, e.g. about 30-35 nucleotides or greater, an algorithm disclosed by Suggs et al, pages 683-693 in Brown, editor, ICN-UCLA Symp. Dev. Biol., Vol. 23 (Academic Press, New York, 1981) may be conveniently employed. Clearly, there are many approaches available to one skilled in the art for designing sets of minimally cross-hybridizing subunits within the scope of the invention. For example, to minimize the effects of different base-stacking energies of terminal nucleotides when subunits are assembled, subunits may be provided that have the same terminal nucleotides. In this way, when subunits are linked, the sum of the base-stacking energies of all the adjoining terminal nucleotides will be the same, thereby reducing or eliminating variability in tag melting temperatures.

A "word" of terminal nucleotides, shown in *italic* below, may also be added to each end of a tag so that a perfect match is always formed between it and a similar terminal "word" on any other tag complement. Such an augmented tag would have the form:

<i>W</i>	<i>W</i> ₁	<i>W</i> ₂	...	<i>W</i> _{k-1}	<i>W</i> _k	<i>W</i>
<i>W'</i>	<i>W</i> ₁ '	<i>W</i> ₂ '	...	<i>W</i> _{k-1} '	<i>W</i> _k '	<i>W'</i>

where the primed *W*'s indicate complements. With ends of tags always forming perfectly matched duplexes, all mismatched words will be internal mismatches thereby reducing the stability of tag-complement duplexes that otherwise would have mismatched words at their ends. It is well known that duplexes with internal mismatches are significantly less stable than duplexes with the same mismatch at a terminus.

A preferred embodiment of minimally cross-hybridizing sets are those whose subunits are made up of three of the four natural nucleotides. As will be discussed more fully below, the absence of one type of nucleotide in the oligonucleotide tags permits target polynucleotides to be loaded onto solid phase supports by use of the 5'→3' exonuclease activity of a DNA polymerase. The following is an exemplary minimally cross-hybridizing set of subunits each comprising four nucleotides selected from the group consisting of A, G, and T:

5

Table II

Word:	w ₁	w ₂	w ₃	w ₄
Sequence:	GATT	TGAT	TAGA	TTTG
Word:	w ₅	w ₆	w ₇	w ₈
Sequence:	GTAA	AGTA	ATGT	AAAG

10 In this set, each member would form a duplex having three mismatched bases with the complement of every other member.

Further exemplary minimally cross-hybridizing sets are listed below in Table III. Clearly, additional sets can be generated by substituting different groups of nucleotides, or by using subsets of known minimally cross-hybridizing sets.

15

Table III

Exemplary Minimally Cross-Hybridizing Sets of 4-mer Subunits

<u>Set 1</u>	<u>Set 2</u>	<u>Set 3</u>	<u>Set 4</u>	<u>Set 5</u>	<u>Set 6</u>
CATT	ACCC	AAAC	AAAG	AACA	AACG
CTAA	AGGG	ACCA	ACCA	ACAC	ACAA
TCAT	CACG	AGGG	AGGC	AGGG	AGGC
ACTA	CCGA	CACG	CACC	CAAG	CAAC
TACA	CGAC	CCGC	CCGG	CCGC	CCGG
TTTC	GAGC	CGAA	CGAA	CGCA	CGCA
ATCT	GCAG	GAGA	GAGA	GAGA	GAGA
AAAC	GGCA	GCAG	GCAC	GCCG	GCCC
	AAAA	GGCC	GGCG	GGAC	GGAG

<u>Set 1</u>	<u>Set 8</u>	<u>Set 9</u>	<u>Set 10</u>	<u>Set 11</u>	<u>Set 12</u>
AAGA	AAGC	AAGG	ACAG	ACCG	ACGA
ACAC	ACAA	ACAA	AACA	AAAA	AAAC
AGCG	AGCG	AGCC	AGGC	AGGC	AGCG
CAAG	CAAG	CAAC	CAAC	CACC	CACA
CCCA	CCCC	CCCG	CCGA	CCGA	CCAG
CGGC	CGGA	CGGA	CGCG	CGAG	CGGC
GACC	GACA	GACA	GAGG	GAGG	GAGG
GCGG	GCGG	GCGC	GCCC	GCAC	GCCC
GGAA	GGAC	GGAG	GGAA	GGCA	GGAA

The oligonucleotide tags of the invention and their complements are conveniently synthesized on an automated DNA synthesizer, e.g. an Applied Biosystems, Inc. (Foster City, California) model 392 or 394 DNA/RNA Synthesizer, using standard chemistries, such as phosphoramidite chemistry, e.g. disclosed in the following references: Beaucage and Iyer, *Tetrahedron*, 48: 2223-2311 (1992); Molko et al, U.S. patent 4,980,460; Koster et al, U.S. patent 4,725,677; Caruthers et al, U.S. patents 4,415,732; 4,458,066; and 4,973,679; and the like. Alternative chemistries, e.g. resulting in non-natural backbone groups, such as phosphorothioate, phosphoramidate, and the like, may also be employed provided that the resulting oligonucleotides are capable of specific hybridization. In some embodiments, tags may comprise naturally occurring nucleotides that permit processing or manipulation by enzymes, while the corresponding tag complements may comprise non-natural nucleotide analogs, such as peptide nucleic acids, or like compounds, that promote the formation of more stable duplexes during sorting.

When microparticles are used as supports, repertoires of oligonucleotide tags and tag complements may be generated by subunit-wise synthesis via "split and mix" techniques, e.g. as disclosed in Shortle et al. International patent application PCT/US93/03418 or Lyttle et al, *Biotechniques*, 19: 274-280 (1995). Briefly, the basic unit of the synthesis is a subunit of the oligonucleotide tag. Preferably, phosphoramidite chemistry is used and 3' phosphoramidite oligonucleotides are prepared for each subunit in a minimally cross-hybridizing set, e.g. for the set first listed above, there would be eight 4-mer 3'-phosphoramidites. Synthesis proceeds as disclosed by Shortle et al or in direct analogy with the techniques employed to generate diverse oligonucleotide libraries using nucleosidic monomers, e.g. as disclosed in Telenius et al, *Genomics*, 13: 718-725 (1992); Welsh et al, *Nucleic Acids Research*, 19: 5275-5279 (1991); Grothues et al, *Nucleic Acids Research*, 21: 1321-1322 (1993); Hartley, European patent application 90304496.4; Lam et al, *Nature*, 354: 82-84 (1991); Zuckerman et al, *Int. J. Pept. Protein Research*, 40: 498-507 (1992); and the like. Generally, these techniques simply call for the application of

mixtures of the activated monomers to the growing oligonucleotide during the coupling steps. Preferably, oligonucleotide tags and tag complements are synthesized on a DNA synthesizer having a number of synthesis chambers which is greater than or equal to the number of different kinds of words used in the construction of the tags.

- 5 That is, preferably there is a synthesis chamber corresponding to each type of word. In this embodiment, words are added nucleotide-by-nucleotide, such that if a word consists of five nucleotides there are five monomer couplings in each synthesis chamber. After a word is completely synthesized, the synthesis supports are removed from the chambers, mixed, and redistributed back to the chambers for the next cycle
10 of word addition. This latter embodiment takes advantage of the high coupling yields of monomer addition, e.g. in phosphoramidite chemistries.

- Double stranded forms of tags may be made by separately synthesizing the complementary strands followed by mixing under conditions that permit duplex formation. Alternatively, double stranded tags may be formed by first synthesizing a
15 single stranded repertoire linked to a known oligonucleotide sequence that serves as a primer binding site. The second strand is then synthesized by combining the single stranded repertoire with a primer and extending with a polymerase. This latter approach is described in Oliphant et al, Gene, 44: 177-183 (1986). Such duplex tags
- may then be inserted into cloning vectors along with target polynucleotides for sorting
20 and manipulation of the target polynucleotide in accordance with the invention.

- When tag complements are employed that are made up of nucleotides that have enhanced binding characteristics, such as PNAs or oligonucleotide N3'→P5' phosphoramidates, sorting can be implemented through the formation of D-loops between tags comprising natural nucleotides and their PNA or phosphoramidate
25 complements, as an alternative to the "stripping" reaction employing the 3'→5' exonuclease activity of a DNA polymerase to render a tag single stranded.

- Oligonucleotide tags of the invention may range in length from 12 to 60 nucleotides or basepairs. Preferably, oligonucleotide tags range in length from 18 to 40 nucleotides or basepairs. More preferably, oligonucleotide tags range in length
30 from 25 to 40 nucleotides or basepairs. In terms of preferred and more preferred numbers of subunits, these ranges may be expressed as follows:

Table IV
Numbers of Subunits in Tags in Preferred Embodiments

35

<u>Monomers in Subunit</u>	<u>Nucleotides in Oligonucleotide Tag</u>		
	(12-60)	(18-40)	(25-40)

3	4-20 subunits	6-13 subunits	8-13 subunits
4	3-15 subunits	4-10 subunits	6-10 subunits
5	2-12 subunits	3-8 subunits	5-8 subunits
6	2-10 subunits	3-6 subunits	4-6 subunits

Most preferably, oligonucleotide tags are single stranded and specific hybridization occurs via Watson-Crick pairing with a tag complement.

Preferably, repertoires of single stranded oligonucleotide tags of the invention
 5 contain at least 100 members; more preferably, repertoires of such tags contain at least 1000 members; and most preferably, repertoires of such tags contain at least 10,000 members.

Triplex Tags

In embodiments where specific hybridization occurs via triplex formation,
 10 coding of tag sequences follows the same principles as for duplex-forming tags; however, there are further constraints on the selection of subunit sequences. Generally, third strand association via Hoogsteen type of binding is most stable along homopyrimidine-homopurine tracks in a double stranded target. Usually, base triplets
 15 form in T-A*T or C-G*C motifs (where "-" indicates Watson-Crick pairing and "*" indicates Hoogsteen type of binding); however, other motifs are also possible. For example, Hoogsteen base pairing permits parallel and antiparallel orientations between the third strand (the Hoogsteen strand) and the purine-rich strand of the duplex to which the third strand binds, depending on conditions and the composition of the strands. There is extensive guidance in the literature for selecting appropriate
 20 sequences, orientation, conditions, nucleoside type (e.g. whether ribose or deoxyribose nucleosides are employed), base modifications (e.g. methylated cytosine, and the like) in order to maximize, or otherwise regulate, triplex stability as desired in particular embodiments, e.g. Roberts et al, Proc. Natl. Acad. Sci., 88: 9397-9401 (1991); Roberts et al, Science, 258: 1463-1466 (1992); Roberts et al, Proc. Natl.
 25 Acad. Sci., 93: 4320-4325 (1996); Distefano et al, Proc. Natl. Acad. Sci., 90: 1179-1183 (1993); Mergny et al, Biochemistry, 30: 9791-9798 (1991); Cheng et al, J. Am. Chem. Soc., 114: 4465-4474 (1992); Beal and Dervan, Nucleic Acids Research, 20: 2773-2776 (1992); Beal and Dervan, J. Am. Chem. Soc., 114: 4976-4982 (1992); Giovannangeli et al, Proc. Natl. Acad. Sci., 89: 8631-8635 (1992); Moser and Dervan,
 30 Science, 238: 645-650 (1987); McShan et al, J. Biol. Chem., 267:5712-5721 (1992); Yoon et al, Proc. Natl. Acad. Sci., 89: 3840-3844 (1992); Blume et al, Nucleic Acids Research, 20: 1777-1784 (1992); Thuong and Helene, Angew. Chem. Int. Ed. Engl.

32: 666-690 (1993); Escude et al, Proc. Natl. Acad. Sci., 93: 4365-4369 (1996); and the like. Conditions for annealing single-stranded or duplex tags to their single-stranded or duplex complements are well known, e.g. Ji et al, Anal. Chem. 65: 1323-1328 (1993); Cantor et al, U.S. patent 5,482,836; and the like. Use of triplex tags has the advantage of not requiring a "stripping" reaction with polymerase to expose the tag for annealing to its complement.

Preferably, oligonucleotide tags of the invention employing triplex hybridization are double stranded DNA and the corresponding tag complements are single stranded. More preferably, 5-methylcytosine is used in place of cytosine in the tag complements in order to broaden the range of pH stability of the triplex formed between a tag and its complement. Preferred conditions for forming triplexes are fully disclosed in the above references. Briefly, hybridization takes place in concentrated salt solution, e.g. 1.0 M NaCl, 1.0 M potassium acetate, or the like, at pH below 5.5 (or 6.5 if 5-methylcytosine is employed). Hybridization temperature depends on the length and composition of the tag; however, for an 18-20-mer tag of longer, hybridization at room temperature is adequate. Washes may be conducted with less concentrated salt solutions, e.g. 10 mM sodium acetate, 100 mM MgCl₂, pH 5.8, at room temperature. Tags may be eluted from their tag complements by incubation in a similar salt solution at pH 9.0.

Minimally cross-hybridizing sets of oligonucleotide tags that form triplexes may be generated by the computer program of Appendix Ic, or similar programs. An exemplary set of double stranded 8-mer words are listed below in capital letters with the corresponding complements in small letters. Each such word differs from each of the other words in the set by three base pairs.

Table V
Exemplary Minimally Cross-Hybridizing
Set of DoubleStranded 8-mer Tags

5' -AAGGAGAG	5' -AAAGGGGA	5' -AGAGAAGA	5' -AGGGGGGG
3' -TTCCTCTC	3' -TTTCCCCT	3' -TCTCTTCT	3' -TCCCCCCC
3' -ttcctctc	3' -tttcccct	3' -tctcttct	3' -tccccccc
5' -AAAAAAA	5' -AAGAGAGA	5' -AGGAAAAG	5' -GAAAGGAG
3' -TTTTTTT	3' -TTCTCTCT	3' -TCCTTTTC	3' -CTTCTCTC
3' -ttttttt	3' -ttctctct	3' -tccttttc	3' -cttctctc
5' -AAAAAGGG	5' -AGAAGAGG	5' -AGGAAGGA	5' -GAAGAAGG
3' -TTTTTCCC	3' -TCTTCTCC	3' -TCCTTCCT	3' -CTTCTTCC
3' -tttttccc	3' -tcttctcc	3' -tccttcct	3' -cttctctc
5' -AAAGGAAG	5' -AGAAGGAA	5' -AGGGGAAA	5' -GAAGAGAA
3' -TTTCCTTC	3' -TCTTCCTT	3' -TCCCCTTT	3' -CTTCTCTT
3' -tttccttc	3' -tcttcctt	3' -tccccttt	3' -cttctctt

5

10

Table VI
Repertoire Size of Various Double Stranded Tags
That Form Triplexes with Their Tag Complements

Oligonucleotide Word Length	Nucleotide Difference between Oligonucleotides of Minimally Cross- Hybridizing Set	Maximal Size of Minimally Cross- Hybridizing Set	Size of Repertoire with Four Words	Size of Repertoire with Five Words
4	2	8	4096	3.2×10^4
6	3	8	4096	3.2×10^4
8	3	16	6.5×10^4	1.05×10^6
10	5	8	4096	
15	5	92		
20	6	765		
20	8	92		
20	10	22		

15 Preferably, repertoires of double stranded oligonucleotide tags of the invention contain at least 10 members; more preferably, repertoires of such tags contain at least 100 members. Preferably, words are between 4 and 8 nucleotides in length for combinatorially synthesized double stranded oligonucleotide tags, and oligonucleotide tags are between 12 and 60 base pairs in length. More preferably, such tags are
20 between 18 and 40 base pairs in length.

Solid Phase Supports

25 Solid phase supports for use with the invention may have a wide variety of forms, including microparticles, beads, and membranes, slides, plates, micromachined chips, and the like. Likewise, solid phase supports of the invention may comprise a

wide variety of compositions, including glass, plastic, silicon, alkanethiolate-derivatized gold, cellulose, low cross-linked and high cross-linked polystyrene, silica gel, polyamide, and the like. Preferably, either a population of discrete particles are employed such that each has a uniform coating, or population, of complementary sequences of the same tag (and no other), or a single or a few supports are employed with spatially discrete regions each containing a uniform coating, or population, of complementary sequences to the same tag (and no other). In the latter embodiment, the area of the regions may vary according to particular applications; usually, the regions range in area from several μm^2 , e.g. 3-5, to several hundred μm^2 , e.g. 100-500. Preferably, such regions are spatially discrete so that signals generated by events, e.g. fluorescent emissions, at adjacent regions can be resolved by the detection system being employed. In some applications, it may be desirable to have regions with uniform coatings of more than one tag complement, e.g. for simultaneous sequence analysis, or for bringing separately tagged molecules into close proximity.

Tag complements may be used with the solid phase support that they are synthesized on, or they may be separately synthesized and attached to a solid phase support for use, e.g. as disclosed by Lund et al, *Nucleic Acids Research*, 16: 10861-10880 (1988); Albretsen et al, *Anal. Biochem.*, 189: 40-50 (1990); Wolf et al, *Nucleic Acids Research*, 15: 2911-2926 (1987); or Ghosh et al, *Nucleic Acids Research*, 15: 5353-5372 (1987). Preferably, tag complements are synthesized on and used with the same solid phase support, which may comprise a variety of forms and include a variety of linking moieties. Such supports may comprise microparticles or arrays, or matrices, of regions where uniform populations of tag complements are synthesized. A wide variety of microparticle supports may be used with the invention, including microparticles made of controlled pore glass (CPG), highly cross-linked polystyrene, acrylic copolymers, cellulose, nylon, dextran, latex, polyacrolein, and the like, disclosed in the following exemplary references: *Meth. Enzymol.*, Section A, pages 11-147, vol. 44 (Academic Press, New York, 1976); U.S. patents 4,678,814; 4,413,070; and 4,046,720; and Pon, Chapter 19, in Agrawal, editor, *Methods in Molecular Biology*, Vol. 20, (Humana Press, Totowa, NJ, 1993). Microparticle supports further include commercially available nucleoside-derivatized CPG and polystyrene beads (e.g. available from Applied Biosystems, Foster City, CA); derivatized magnetic beads; polystyrene grafted with polyethylene glycol (e.g., TentaGelTM, Rapp Polymere, Tübingen Germany); and the like. Selection of the support characteristics, such as material, porosity, size, shape, and the like, and the type of linking moiety employed depends on the conditions under which the tags are used. For example, in applications involving successive processing with enzymes, supports and linkers that minimize steric hindrance of the enzymes and that facilitate

access to substrate are preferred. Other important factors to be considered in selecting the most appropriate microparticle support include size uniformity, efficiency as a synthesis support, degree to which surface area known, and optical properties, e.g. as explain more fully below, clear smooth beads provide instrumental advantages
5 when handling large numbers of beads on a surface.

Exemplary linking moieties for attaching and/or synthesizing tags on microparticle surfaces are disclosed in Pon et al, *Biotechniques*, 6:768-775 (1988); Webb, U.S. patent 4,659,774; Barany et al, International patent application PCT/US91/06103; Brown et al, *J. Chem. Soc. Commun.*, 1989: 891-893; Damha et
10 al, *Nucleic Acids Research*, 18: 3813-3821 (1990); Beattie et al, *Clinical Chemistry*, 39: 719-722 (1993); Maskos and Southern, *Nucleic Acids Research*, 20: 1679-1684 (1992); and the like.

As mentioned above, tag complements may also be synthesized on a single (or a few) solid phase support to form an array of regions uniformly coated with tag
15 complements. That is, within each region in such an array the same tag complement is synthesized. Techniques for synthesizing such arrays are disclosed in McGall et al, International application PCT/US93/03767; Pease et al, *Proc. Natl. Acad. Sci.*, 91: 5022-5026 (1994); Southern and Maskos, International application PCT/GB89/01114; Maskos and Southern (cited above); Southern et al, *Genomics*, 13:
20 1008-1017 (1992); and Maskos and Southern, *Nucleic Acids Research*, 21: 4663-4669 (1993).

Preferably, the invention is implemented with microparticles or beads uniformly coated with complements of the same tag sequence. Microparticle supports and methods of covalently or noncovalently linking oligonucleotides to their surfaces
25 are well known, as exemplified by the following references: Beaucage and Iyer (cited above); Gait, editor, *Oligonucleotide Synthesis: A Practical Approach* (IRL Press, Oxford, 1984); and the references cited above. Generally, the size and shape of a microparticle is not critical; however, microparticles in the size range of a few, e.g. 1-2, to several hundred, e.g. 200-1000 μm diameter are preferable, as they facilitate the
30 construction and manipulation of large repertoires of oligonucleotide tags with minimal reagent and sample usage.

In some preferred applications, commercially available controlled-pore glass (CPG) or polystyrene supports are employed as solid phase supports in the invention. Such supports come available with base-labile linkers and initial nucleosides attached.
35 e.g. Applied Biosystems (Foster City, CA). Preferably, microparticles having pore size between 500 and 1000 angstroms are employed.

In other preferred applications, non-porous microparticles are employed for their optical properties, which may be advantageously used when tracking large

numbers of microparticles on planar supports, such as a microscope slide.

Particularly preferred non-porous microparticles are the glycidal methacrylate (GMA) beads available from Bangs Laboratories (Carmel, IN). Such microparticles are useful in a variety of sizes and derivatized with a variety of linkage groups for synthesizing tags or tag complements. Preferably, for massively parallel manipulations of tagged microparticles, 5 μ m diameter GMA beads are employed.

Attaching Tags to Polynucleotides

For Sorting onto Solid Phase Supports

An important aspect of the invention is the sorting and attachment of a populations of polynucleotides, e.g. from a cDNA library, to microparticles or to separate regions on a solid phase support such that each microparticle or region has substantially only one kind of polynucleotide attached. This objective is accomplished by insuring that substantially all different polynucleotides have different tags attached. This condition, in turn, is brought about by taking a sample of the full ensemble of tag-polynucleotide conjugates for analysis. (It is acceptable that identical polynucleotides have different tags, as it merely results in the same polynucleotide being operated on or analyzed twice in two different locations.) Such sampling can be carried out either overtly--for example, by taking a small volume from a larger mixture--after the tags have been attached to the polynucleotides, it can be carried out inherently as a secondary effect of the techniques used to process the polynucleotides and tags, or sampling can be carried out both overtly and as an inherent part of processing steps.

Preferably, in constructing a cDNA library where substantially all different cDNAs have different tags, a tag repertoire is employed whose complexity, or number of distinct tags, greatly exceeds the total number of mRNAs extracted from a cell or tissue sample. Preferably, the complexity of the tag repertoire is at least 10 times that of the polynucleotide population; and more preferably, the complexity of the tag repertoire is at least 100 times that of the polynucleotide population. Below, a protocol is disclosed for cDNA library construction using a primer mixture that contains a full repertoire of exemplary 9-word tags. Such a mixture of tag-containing primers has a complexity of 8^9 , or about 1.34×10^8 . As indicated by Winslow et al, Nucleic Acids Research, 19: 3251-3253 (1991), mRNA for library construction can be extracted from as few as 10-100 mammalian cells. Since a single mammalian cell contains about 5×10^5 copies of mRNA molecules of about 3.4×10^4 different kinds,

by standard techniques one can isolate the mRNA from about 100 cells, or (theoretically) about 5×10^7 mRNA molecules. Comparing this number to the complexity of the primer mixture shows that without any additional steps, and even assuming that mRNAs are converted into cDNAs with perfect efficiency (1% efficiency or less is more accurate), the cDNA library construction protocol results in a population containing no more than 37% of the total number of different tags. That is, without any overt sampling step at all, the protocol inherently generates a sample that comprises 37%, or less, of the tag repertoire. The probability of obtaining a double under these conditions is about 5%, which is within the preferred range. With mRNA from 10 cells, the fraction of the tag repertoire sampled is reduced to only 3.7%, even assuming that all the processing steps take place at 100% efficiency. In fact, the efficiencies of the processing steps for constructing cDNA libraries are very low, a "rule of thumb" being that good library should contain about 10^8 cDNA clones from mRNA extracted from 10^6 mammalian cells.

Use of larger amounts of mRNA in the above protocol, or for larger amounts of polynucleotides in general, where the number of such molecules exceeds the complexity of the tag repertoire, a tag-polynucleotide conjugate mixture potentially contains every possible pairing of tags and types of mRNA or polynucleotide. In such cases, overt sampling may be implemented by removing a sample volume after a serial dilution of the starting mixture of tag-polynucleotide conjugates. The amount of dilution required depends on the amount of starting material and the efficiencies of the processing steps, which are readily estimated.

If mRNA were extracted from 10^6 cells (which would correspond to about 0.5 μg of poly(A)⁺ RNA), and if primers were present in about 10-100 fold concentration excess--as is called for in a typical protocol, e.g. Sambrook et al, Molecular Cloning, Second Edition, page 8.61 [10 μL 1.8 kb mRNA at 1 mg/mL equals about 1.68×10^{-11} moles and 10 μL 18-mer primer at 1 mg/mL equals about 1.68×10^{-9} moles], then the total number of tag-polynucleotide conjugates in a cDNA library would simply be equal to or less than the starting number of mRNAs, or about 5×10^{11} vectors containing tag-polynucleotide conjugates--again this assumes that each step in cDNA construction--first strand synthesis, second strand synthesis, ligation into a vector--occurs with perfect efficiency, which is a very conservative estimate. The actual number is significantly less.

If a sample of n tag-polynucleotide conjugates are randomly drawn from a reaction mixture--as could be effected by taking a sample volume, the probability of drawing conjugates having the same tag is described by the Poisson distribution, $P(r) = e^{-\lambda} (\lambda)^r / r!$, where r is the number of conjugates having the same tag and $\lambda = np$, where p is the probability of a given tag being selected. If $n = 10^6$ and $p = 1/(1.34 \times$

10⁸), then $\lambda = .00746$ and $P(2) = 2.76 \times 10^{-5}$. Thus, a sample of one million molecules gives rise to an expected number of doubles well within the preferred range. Such a sample is readily obtained as follows: Assume that the 5×10^{11} mRNAs are perfectly converted into 5×10^{11} vectors with tag-cDNA conjugates as inserts and that the 5×10^{11} vectors are in a reaction solution having a volume of 100 μ l. Four 10-fold serial dilutions may be carried out by transferring 10 μ l from the original solution into a vessel containing 90 μ l of an appropriate buffer, such as TE. This process may be repeated for three additional dilutions to obtain a 100 μ l solution containing 5×10^5 vector molecules per μ l. A 2 μ l aliquot from this solution yields 10^6 vectors containing tag-cDNA conjugates as inserts. This sample is then amplified by straight forward transformation of a competent host cell followed by culturing.

Of course, as mentioned above, no step in the above process proceeds with perfect efficiency. In particular, when vectors are employed to amplify a sample of tag-polynucleotide conjugates, the step of transforming a host is very inefficient. Usually, no more than 1% of the vectors are taken up by the host and replicated. Thus, for such a method of amplification, even fewer dilutions would be required to obtain a sample of 10^6 conjugates.

A repertoire of oligonucleotide tags can be conjugated to a population of polynucleotides in a number of ways, including direct enzymatic ligation, amplification, e.g. via PCR, using primers containing the tag sequences, and the like. The initial ligating step produces a very large population of tag-polynucleotide conjugates such that a single tag is generally attached to many different polynucleotides. However, as noted above, by taking a sufficiently small sample of the conjugates, the probability of obtaining "doubles," i.e. the same tag on two different polynucleotides, can be made negligible. Generally, the larger the sample the greater the probability of obtaining a double. Thus, a design trade-off exists between selecting a large sample of tag-polynucleotide conjugates--which, for example, ensures adequate coverage of a target polynucleotide in a shotgun sequencing operation or adequate representation of a rapidly changing mRNA pool, and selecting a small sample which ensures that a minimal number of doubles will be present. In most embodiments, the presence of doubles merely adds an additional source of noise or, in the case of sequencing, a minor complication in scanning and signal processing, as microparticles giving multiple fluorescent signals can simply be ignored.

As used herein, the term "substantially all" in reference to attaching tags to molecules, especially polynucleotides, is meant to reflect the statistical nature of the sampling procedure employed to obtain a population of tag-molecule conjugates essentially free of doubles. The meaning of substantially all in terms of actual

percentages of tag-molecule conjugates depends on how the tags are being employed. Preferably, for nucleic acid sequencing, substantially all means that at least eighty percent of the polynucleotides have unique tags attached. More preferably, it means that at least ninety percent of the polynucleotides have unique tags attached. Still
 5 more preferably, it means that at least ninety-five percent of the polynucleotides have unique tags attached. And, most preferably, it means that at least ninety-nine percent of the polynucleotides have unique tags attached.

Preferably, when the population of polynucleotides consists of messenger RNA (mRNA), oligonucleotides tags may be attached by reverse transcribing the
 10 mRNA with a set of primers preferably containing complements of tag sequences. An exemplary set of such primers could have the following sequence (SEQ ID NO: 1):

5' -mRNA- [A]_n -3'
 15 [T]₁₉GG[W,W,W,C]₉ACCAGCTGATC-5'-biotin

where "[W,W,W,C]₉" represents the sequence of an oligonucleotide tag of nine
 - subunits of four nucleotides each and "[W,W,W,C]" represents the subunit sequences
 20 listed above, i.e. "W" represents T or A. The underlined sequences identify an optional restriction endonuclease site that can be used to release the polynucleotide from attachment to a solid phase support via the biotin, if one is employed. For the above primer, the complement attached to a microparticle could have the form:

25 5' -[G,W,W,W]₉TGG-linker-microparticle

After reverse transcription, the mRNA is removed, e.g. by RNase H digestion. and the second strand of the cDNA is synthesized using, for example, a primer of the following form (SEQ ID NO: 2):

30 5' -NRRGATCYNNN-3'

where N is any one of A, T, G, or C; R is a purine-containing nucleotide, and Y is a pyrimidine-containing nucleotide. This particular primer creates a Bst Y1 restriction
 35 site in the resulting double stranded DNA which, together with the Sal I site, facilitates cloning into a vector with, for example, Bam HI and Xho I sites. After Bst Y1 and Sal I digestion, the exemplary conjugate would have the form:

5'-RCGACCA[C,W,W,W]9GG[T]19- cDNA -NNNR
GGT[G,W,W,W]9CC[A]19- rDNA -NNNYCTAG-5'

The polynucleotide-tag conjugates may then be manipulated using standard molecular biology techniques. For example, the above conjugate--which is actually a mixture--may be inserted into commercially available cloning vectors, e.g. Stratagene Cloning System (La Jolla, CA); transfected into a host, such as a commercially available host bacteria; which is then cultured to increase the number of conjugates. The cloning vectors may then be isolated using standard techniques, e.g. Sambrook et al, Molecular Cloning, Second Edition (Cold Spring Harbor Laboratory, New York, 1989). Alternatively, appropriate adaptors and primers may be employed so that the conjugate population can be increased by PCR.

Preferably, when the ligase-based method of sequencing is employed, the Bst YI and Sal I digested fragments are cloned into a Bam HI-/Xho I-digested vector having the following single-copy restriction sites (SEQ ID NO: 3):

5'-GAGGATGCCTTTATGGATCCACTCGAGATCCCAATCCA-3'
FokI BamHI XhoI

This adds the Fok I site which will allow initiation of the sequencing process discussed more fully below.

Tags can be conjugated to cDNAs of existing libraries by standard cloning methods. cDNAs are excised from their existing vector, isolated, and then ligated into a vector containing a repertoire of tags. Preferably, the tag-containing vector is linearized by cleaving with two restriction enzymes so that the excised cDNAs can be ligated in a predetermined orientation. The concentration of the linearized tag-containing vector is in substantial excess over that of the cDNA inserts so that ligation provides an inherent sampling of tags.

A general method for exposing the single stranded tag after amplification involves digesting a target polynucleotide-containing conjugate with the 5'→3' exonuclease activity of T4 DNA polymerase, or a like enzyme. When used in the presence of a single deoxynucleoside triphosphate, such a polymerase will cleave nucleotides from 3' recessed ends present on the non-template strand of a double stranded fragment until a complement of the single deoxynucleoside triphosphate is reached on the template strand. When such a nucleotide is reached the 5'→3' digestion effectively ceases, as the polymerase's extension activity adds nucleotides at a higher rate than the excision activity removes nucleotides. Consequently, single

stranded tags constructed with three nucleotides are readily prepared for loading onto solid phase supports.

The technique may also be used to preferentially methylate interior Fok I sites of a target polynucleotide while leaving a single Fok I site at the terminus of the polynucleotide unmethylated. First, the terminal Fok I site is rendered single stranded using a polymerase with deoxycytidine triphosphate. The double stranded portion of the fragment is then methylated, after which the single stranded terminus is filled in with a DNA polymerase in the presence of all four nucleoside triphosphates, thereby regenerating the Fok I site. Clearly, this procedure can be generalized to endonucleases other than Fok I.

After the oligonucleotide tags are prepared for specific hybridization, e.g. by rendering them single stranded as described above, the polynucleotides are mixed with microparticles containing the complementary sequences of the tags under conditions that favor the formation of perfectly matched duplexes between the tags and their complements. There is extensive guidance in the literature for creating these conditions. Exemplary references providing such guidance include Wetmur, *Critical Reviews in Biochemistry and Molecular Biology*, 26: 227-259 (1991); Sambrook et al, *Molecular Cloning: A Laboratory Manual*, 2nd Edition (Cold Spring Harbor Laboratory, New York, 1989); and the like. Preferably, the hybridization conditions are sufficiently stringent so that only perfectly matched sequences form stable duplexes. Under such conditions the polynucleotides specifically hybridized through their tags may be ligated to the complementary sequences attached to the microparticles. Finally, the microparticles are washed to remove polynucleotides with unligated and/or mismatched tags.

When CPG microparticles conventionally employed as synthesis supports are used, the density of tag complements on the microparticle surface is typically greater than that necessary for some sequencing operations. That is, in sequencing approaches that require successive treatment of the attached polynucleotides with a variety of enzymes, densely spaced polynucleotides may tend to inhibit access of the relatively bulky enzymes to the polynucleotides. In such cases, the polynucleotides are preferably mixed with the microparticles so that tag complements are present in significant excess, e.g. from 10:1 to 100:1, or greater, over the polynucleotides. This ensures that the density of polynucleotides on the microparticle surface will not be so high as to inhibit enzyme access. Preferably, the average inter-polynucleotide spacing on the microparticle surface is on the order of 30-100 nm. Guidance in selecting ratios for standard CPG supports and Ballotini beads (a type of solid glass support) is found in Maskos and Southern, *Nucleic Acids Research*, 20: 1679-1684 (1992). Preferably, for sequencing applications, standard CPG beads of diameter in the range

of 20-50 μm are loaded with about 10^5 polynucleotides, and GMA beads of diameter in the range of 5-10 μm are loaded with a few tens of thousand of polynucleotides, e.g. 4×10^4 to 6×10^4 .

In the preferred embodiment, tag complements are synthesized on
 5 microparticles combinatorially; thus, at the end of the synthesis, one obtains a complex mixture of microparticles from which a sample is taken for loading tagged polynucleotides. The size of the sample of microparticles will depend on several factors, including the size of the repertoire of tag complements, the nature of the apparatus for used for observing loaded microparticles--e.g. its capacity, the tolerance
 10 for multiple copies of microparticles with the same tag complement (i.e. "bead doubles"), and the like. The following table provide guidance regarding microparticle sample size, microparticle diameter, and the approximate physical dimensions of a packed array of microparticles of various diameters.

15

Microparticle diameter	5 μm	10 μm	20 μm	40 μm
Max. no. polynucleotides loaded at 1 per 10^5 sq. angstrom		3×10^5	1.26×10^6	5×10^6
Approx. area of monolayer of 10^6 microparticles	.45 x .45 cm	1 x 1 cm	2 x 2 cm	4 x 4 cm

20 The probability that the sample of microparticles contains a given tag complement or is present in multiple copies is described by the Poisson distribution, as indicated in the following table.

25

Table VII

Number of microparticles in sample (as fraction of repertoire size), m	Fraction of repertoire of tag complements present in sample, $1-e^{-m}$	Fraction of microparticles in sample with unique tag complement attached, $m(e^{-m})/2$	Fraction of microparticles in sample carrying same tag complement as one other microparticle in sample ("bead doubles"), $m^2(e^{-m})/2$
1.000	0.63	0.37	0.18
.693	0.50	0.35	0.12
.405	0.33	0.27	0.05
.285	0.25	0.21	0.03
.223	0.20	0.18	0.02
.105	0.10	0.09	0.005
.010	0.01	0.01	

High Specificity Sorting and Panning

5 The kinetics of sorting depends on the rate of hybridization of oligonucleotide tags to their tag complements which, in turn, depends on the complexity of the tags in the hybridization reaction. Thus, a trade off exists between sorting rate and tag complexity, such that an increase in sorting rate may be achieved at the cost of reducing the complexity of the tags involved in the hybridization reaction. As explained below, the effects of this trade off may be ameliorated by "panning."

10 Specificity of the hybridizations may be increased by taking a sufficiently small sample so that both a high percentage of tags in the sample are unique and the nearest neighbors of substantially all the tags in a sample differ by at least two words. This latter condition may be met by taking a sample that contains a number of tag-
15 polynucleotide conjugates that is about 0.1 percent or less of the size of the repertoire being employed. For example, if tags are constructed with eight words selected from Table II, a repertoire of 8^8 , or about 1.67×10^7 , tags and tag complements are produced. In a library of tag-cDNA conjugates as described above, a 0.1 percent sample means that about 16,700 different tags are present. If this were loaded directly
20 onto a repertoire-equivalent of microparticles, or in this example a sample of 1.67×10^7 microparticles, then only a sparse subset of the sampled microparticles would be loaded. The density of loaded microparticles can be increase--for example, for more efficient sequencing--by undertaking a "panning" step in which the sampled tag-cDNA conjugates are used to separate loaded microparticles from unloaded
25 microparticles. Thus, in the example above, even though a "0.1 percent" sample

contains only 16,700 cDNAs, the sampling and panning steps may be repeated until as many loaded microparticles as desired are accumulated.

A panning step may be implemented by providing a sample of tag-cDNA conjugates each of which contains a capture moiety at an end opposite, or distal to, the oligonucleotide tag. Preferably, the capture moiety is of a type which can be released from the tag-cDNA conjugates, so that the tag-cDNA conjugates can be sequenced with a single-base sequencing method. Such moieties may comprise biotin, digoxigenin, or like ligands, a triplex binding region, or the like. Preferably, such a capture moiety comprises a biotin component. Biotin may be attached to tag-cDNA conjugates by a number of standard techniques. If appropriate adapters containing PCR primer binding sites are attached to tag-cDNA conjugates, biotin may be attached by using a biotinylated primer in an amplification after sampling. Alternatively, if the tag-cDNA conjugates are inserts of cloning vectors, biotin may be attached after excising the tag-cDNA conjugates by digestion with an appropriate restriction enzyme followed by isolation and filling in a protruding strand distal to the tags with a DNA polymerase in the presence of biotinylated uridine triphosphate.

After a tag-cDNA conjugate is captured, it may be released from the biotin moiety in a number of ways, such as by a chemical linkage that is cleaved by reduction, e.g. Herman et al, *Anal. Biochem.*, 156: 48-55 (1986), or that is cleaved photochemically, e.g. Olejnik et al, *Nucleic Acids Research*, 24: 361-366 (1996), or that is cleaved enzymatically by introducing a restriction site in the PCR primer. The latter embodiment can be exemplified by considering the library of tag-polynucleotide conjugates described above:

5' - RCGACCA [C, W, W, W] ₉ GG [T] ₁₉ - cDNA - NNNR
GGT [G, W, W, W] ₉ CC [A] ₁₉ - rDNA - NNNYCTAG - 5'

The following adapters may be ligated to the ends of these fragments to permit amplification by PCR:

5' - XXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXYCAT

Right Adapter

GATCZZACTAGTZZZZZZZZZZZZ-3'
ZZTGATCAZZZZZZZZZZZZ

Left Adapter

ZZ**TGATCA**ZZZZZZZZZZZZ-5'-biotin

5

Left Primer

where "ACTAGT" is a Spe I recognition site (which leaves a staggered cleavage ready for single base sequencing), and the X's and Z's are nucleotides selected so that the annealing and dissociation temperatures of the respective primers are approximately the same. After ligation of the adapters and amplification by PCR using the biotinylated primer, the tags of the conjugates are rendered single stranded by the exonuclease activity of T4 DNA polymerase and conjugates are combined with a sample of microparticles, e.g. a repertoire equivalent, with tag complements attached. After annealing under stringent conditions (to minimize mis-attachment of tags), the conjugates are preferably ligated to their tag complements and the loaded microparticles are separated from the unloaded microparticles by capture with avidinated magnetic beads, or like capture technique.

Returning to the example, this process results in the accumulation of about 10,500 (=16,700 x .63) loaded microparticles with different tags, which may be released from the magnetic beads by cleavage with Spe I. By repeating this process 40-50 times with new samples of microparticles and tag-cDNA conjugates, $4-5 \times 10^5$ cDNAs can be accumulated by pooling the released microparticles. The pooled microparticles may then be simultaneously sequenced by a single-base sequencing technique.

Determining how many times to repeat the sampling and panning steps--or more generally, determining how many cDNAs to analyze, depends on one's objective. If the objective is to monitor the changes in abundance of relatively common sequences, e.g. making up 5% or more of a population, then relatively small samples, i.e. a small fraction of the total population size, may allow statistically significant estimates of relative abundances. On the other hand, if one seeks to monitor the abundances of rare sequences, e.g. making up 0.1% or less of a population, then large samples are required. Generally, there is a direct relationship between sample size and the reliability of the estimates of relative abundances based on the sample. There is extensive guidance in the literature on determining appropriate sample sizes for making reliable statistical estimates, e.g. Koller et al, *Nucleic Acids Research*, 23:185-191 (1994); Good, *Biometrika*, 40: 16-264 (1953); Bunge et al, *J. Am. Stat. Assoc.*, 88: 364-373 (1993); and the like. Preferably, for

monitoring changes in gene expression based on the analysis of a series of cDNA libraries containing 10^5 to 10^8 independent clones of $3.0-3.5 \times 10^4$ different sequences, a sample of at least 10^4 sequences are accumulated for analysis of each library. More preferably, a sample of at least 10^5 sequences are accumulated for the analysis of each library; and most preferably, a sample of at least 5×10^5 sequences are accumulated for the analysis of each library. Alternatively, the number of sequences sampled is preferably sufficient to estimate the relative abundance of a sequence present at a frequency within the range of 0.1% to 5% with a 95% confidence limit no larger than 0.1% of the population size.

Single Base DNA Sequencing

The present invention can be employed with conventional methods of DNA sequencing, e.g. as disclosed by Hultman et al, Nucleic Acids Research, 17: 4937-4946 (1989). However, for parallel, or simultaneous, sequencing of multiple polynucleotides, a DNA sequencing methodology is preferred that requires neither electrophoretic separation of closely sized DNA fragments nor analysis of cleaved nucleotides by a separate analytical procedure, as in peptide sequencing. Preferably, the methodology permits the stepwise identification of nucleotides, usually one at a time, in a sequence through successive cycles of treatment and detection. Such methodologies are referred to herein as "single base" sequencing methods. Single base approaches are disclosed in the following references: Cheeseman, U.S. patent 5,302,509; Tsien et al, International application WO 91/06678; Rosenthal et al, International application WO 93/21340; Canard et al, Gene, 148: 1-6 (1994); and Metzker et al, Nucleic Acids Research, 22: 4259-4267 (1994).

A "single base" method of DNA sequencing which is suitable for use with the present invention and which requires no electrophoretic separation of DNA fragments is described in International application PCT/US95/03678. Briefly, the method comprises the following steps: (a) ligating a probe to an end of the polynucleotide having a protruding strand to form a ligated complex, the probe having a complementary protruding strand to that of the polynucleotide and the probe having a nuclease recognition site; (b) removing unligated probe from the ligated complex; (c) identifying one or more nucleotides in the protruding strand of the polynucleotide by the identity of the ligated probe; (d) cleaving the ligated complex with a nuclease; and (e) repeating steps (a) through (d) until the nucleotide sequence of the polynucleotide, or a portion thereof, is determined.

A single signal generating moiety, such as a single fluorescent dye, may be employed when sequencing several different target polynucleotides attached to different spatially addressable solid phase supports, such as fixed microparticles, in a

parallel sequencing operation. This may be accomplished by providing four sets of probes that are applied sequentially to the plurality of target polynucleotides on the different microparticles. An exemplary set of such probes are shown below:

5

Set 1	Set 2	Set 3	Set 4
ANNNN...NN N...NNTT...T*	dANNNN...NN d N...NNTT...T	dANNNN...NN N...NNTT...T	dANNNN...NN N...NNTT...T
dCNNNN...NN N...NNTT...T	CNNNN...NN N...NNTT...T*	dCNNNN...NN N...NNTT...T	dCNNNN...NN N...NNTT...T
dGNNNN...NN N...NNTT...T	dGNNNN...NN N...NNTT...T	GNNNN...NN N...NNTT...T*	dGNNNN...NN N...NNTT...T
dTNNNN...NN N...NNTT...T	dTNNNN...NN N...NNTT...T	dTNNNN...NN N...NNTT...T	TNNNN...NN N...NNTT...T*

where each of the listed probes represents a mixture of $4^3=64$ oligonucleotides such that the identity of the 3' terminal nucleotide of the top strand is fixed and the other positions in the protruding strand are filled by every 3-mer permutation of nucleotides, or complexity reducing analogs. The listed probes are also shown with a single stranded poly-T tail with a signal generating moiety attached to the terminal thymidine, shown as "T*". The "d" on the unlabeled probes designates a ligation-blocking moiety or absence of 3'-hydroxyl, which prevents unlabeled probes from being ligated. Preferably, such 3'-terminal nucleotides are dideoxynucleotides. In this embodiment, the probes of set 1 are first applied to the plurality of target polynucleotides and treated with a ligase so that target polynucleotides having a thymidine complementary to the 3' terminal adenosine of the labeled probes are ligated. The unlabeled probes are simultaneously applied to minimize inappropriate ligations. The locations of the target polynucleotides that form ligated complexes with probes terminating in "A" are identified by the signal generated by the label carried on the probe. After washing and cleavage, the probes of set 2 are applied. In this case, target polynucleotides forming ligated complexes with probes terminating in "C" are identified by location. Similarly, the probes of sets 3 and 4 are applied and locations of positive signals identified. This process of sequentially applying the four sets of probes continues until the desired number of nucleotides are identified on the target polynucleotides. Clearly, one of ordinary skill could construct similar sets of probes that could have many variations, such as having protruding strands of different lengths, different moieties to block ligation of unlabeled probes, different means for labeling probes, and the like.

Apparatus for Sequencing Populations of Polynucleotides

An objective of the invention is to sort identical molecules, particularly polynucleotides, onto the surfaces of microparticles by the specific hybridization of tags and their complements. Once such sorting has taken place, the presence of the molecules or operations performed on them can be detected in a number of ways depending on the nature of the tagged molecule, whether microparticles are detected separately or in "batches," whether repeated measurements are desired, and the like. Typically, the sorted molecules are exposed to ligands for binding, e.g. in drug development, or are subjected chemical or enzymatic processes, e.g. in polynucleotide sequencing. In both of these uses it is often desirable to simultaneously observe signals corresponding to such events or processes on large numbers of microparticles. Microparticles carrying sorted molecules (referred to herein as "loaded" microparticles) lend themselves to such large scale parallel operations, e.g. as demonstrated by Lam et al (cited above).

Preferably, whenever light-generating signals, e.g. chemiluminescent, fluorescent, or the like, are employed to detect events or processes, loaded microparticles are spread on a planar substrate, e.g. a glass slide, for examination with a scanning system, such as described in International patent applications PCT/US91/09217, PCT/NL90/00081, and PCT/US95/01886. The scanning system should be able to reproducibly scan the substrate and to define the positions of each microparticle in a predetermined region by way of a coordinate system. In polynucleotide sequencing applications, it is important that the positional identification of microparticles be repeatable in successive scan steps.

Such scanning systems may be constructed from commercially available components, e.g. x-y translation table controlled by a digital computer used with a detection system comprising one or more photomultiplier tubes, or alternatively, a CCD array, and appropriate optics, e.g. for exciting, collecting, and sorting fluorescent signals. In some embodiments a confocal optical system may be desirable. An exemplary scanning system suitable for use in four-color sequencing is illustrated diagrammatically in Figure 5. Substrate 300, e.g. a microscope slide with fixed microparticles, is placed on x-y translation table 302, which is connected to and controlled by an appropriately programmed digital computer 304 which may be any of a variety of commercially available personal computers, e.g. 486-based machines or PowerPC model 7100 or 8100 available from Apple Computer (Cupertino, CA). Computer software for table translation and data collection functions can be provided by commercially available laboratory software, such as Lab Windows, available from National Instruments.

Substrate 300 and table 302 are operationally associated with microscope 306 having one or more objective lenses 308 which are capable of collecting and delivering light to microparticles fixed to substrate 300. Excitation beam 310 from light source 312, which is preferably a laser, is directed to beam splitter 314, e.g. a dichroic mirror, which re-directs the beam through microscope 306 and objective lens 308 which, in turn, focuses the beam onto substrate 300. Lens 308 collects fluorescence 316 emitted from the microparticles and directs it through beam splitter 314 to signal distribution optics 318 which, in turn, directs fluorescence to one or more suitable opto-electronic devices for converting some fluorescence characteristic, e.g. intensity, lifetime, or the like, to an electrical signal. Signal distribution optics 318 may comprise a variety of components standard in the art, such as bandpass filters, fiber optics, rotating mirrors, fixed position mirrors and lenses, diffraction gratings, and the like. As illustrated in Figure 2, signal distribution optics 318 directs fluorescence 316 to four separate photomultiplier tubes, 330, 332, 334, and 336, whose output is then directed to pre-amps and photon counters 350, 352, 354, and 356. The output of the photon counters is collected by computer 304, where it can be stored, analyzed, and viewed on video 360. Alternatively, signal distribution optics 318 could be a diffraction grating which directs fluorescent signal 318 onto a CCD array.

The stability and reproducibility of the positional localization in scanning will determine, to a large extent, the resolution for separating closely spaced microparticles. Preferably, the scanning systems should be capable of resolving closely spaced microparticles, e.g. separated by a particle diameter or less. Thus, for most applications, e.g. using CPG microparticles, the scanning system should at least have the capability of resolving objects on the order of 10-100 μm . Even higher resolution may be desirable in some embodiments, but with increase resolution, the time required to fully scan a substrate will increase; thus, in some embodiments a compromise may have to be made between speed and resolution. Increases in scanning time can be achieved by a system which only scans positions where microparticles are known to be located, e.g. from an initial full scan. Preferably, microparticle size and scanning system resolution are selected to permit resolution of fluorescently labeled microparticles randomly disposed on a plane at a density between about ten thousand to one hundred thousand microparticles per cm^2 .

In sequencing applications, loaded microparticles can be fixed to the surface of a substrate in variety of ways. The fixation should be strong enough to allow the microparticles to undergo successive cycles of reagent exposure and washing without significant loss. When the substrate is glass, its surface may be derivatized with an alkylamino linker using commercially available reagents, e.g. Pierce Chemical, which

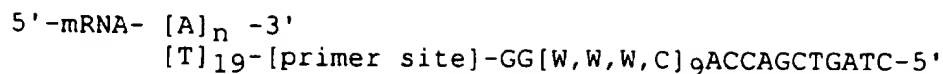
in turn may be cross-linked to avidin, again using conventional chemistries, to form an avidinated surface. Biotin moieties can be introduced to the loaded microparticles in a number of ways. For example, a fraction, e.g. 10-15 percent, of the cloning vectors used to attach tags to polynucleotides are engineered to contain a unique
5 restriction site (providing sticky ends on digestion) immediately adjacent to the polynucleotide insert at an end of the polynucleotide opposite of the tag. The site is excised with the polynucleotide and tag for loading onto microparticles. After loading, about 10-15 percent of the loaded polynucleotides will possess the unique restriction site distal from the microparticle surface. After digestion with the
10 associated restriction endonuclease, an appropriate double stranded adaptor containing a biotin moiety is ligated to the sticky end. The resulting microparticles are then spread on the avidinated glass surface where they become fixed via the biotin-avidin linkages.

Alternatively and preferably when sequencing by ligation is employed, in the
15 initial ligation step a mixture of probes is applied to the loaded microparticle: a fraction of the probes contain a type II restriction recognition site, as required by the sequencing method, and a fraction of the probes have no such recognition site, but instead contain a biotin moiety at its non-ligating end. Preferably, the mixture
- comprises about 10-15 percent of the biotinylated probe.

20 In still another alternative, when DNA-loaded microparticles are applied to a glass substrate, the DNA may nonspecifically adsorb to the glass surface upon several hours, e.g. 24 hours, incubation to create a bond sufficiently strong to permit repeated exposures to reagents and washes without significant loss of microparticles. Preferably, such a glass substrate is a flow cell, which may comprise a channel etched
25 in a glass slide. Preferably, such a channel is closed so that fluids may be pumped through it and has a depth sufficiently close to the diameter of the microparticles so that a monolayer of microparticles is trapped within a defined observation region.

Identification of Novel Polynucleotides in cDNA Libraries

30 Novel polynucleotides in a cDNA library can be identified by constructing a library of cDNA molecules attached to microparticles, as described above. A large fraction of the library, or even the entire library, can then be partially sequenced in parallel. After isolation of mRNA, and perhaps normalization of the population as
35 taught by Soares et al, Proc. Natl. Acad. Sci., 91: 9228-9232 (1994), or like references, the following primer may be hybridized to the polyA tails for first strand synthesis with a reverse transcriptase using conventional protocols (SEQ ID NO: 1):



where [W,W,W,C]₉ represents a tag as described above, "ACCAGCTGATC" is an optional sequence forming a restriction site in double stranded form, and "primer site" is a sequence common to all members of the library that is later used as a primer binding site for amplifying polynucleotides of interest by PCR.

After reverse transcription and second strand synthesis by conventional techniques, the double stranded fragments are inserted into a cloning vector as described above and amplified. The amplified library is then sampled and the sample amplified. The cloning vectors from the amplified sample are isolated, and the tagged cDNA fragments excised and purified. After rendering the tag single stranded with a polymerase as described above, the fragments are methylated and sorted onto microparticles in accordance with the invention. Preferably, as described above, the cloning vector is constructed so that the tagged cDNAs can be excised with an endonuclease, such as Fok I, that will allow immediate sequencing by the preferred single base method after sorting and ligation to microparticles.

Stepwise sequencing is then carried out simultaneously on the whole library, or one or more large fractions of the library, in accordance with the invention until a sufficient number of nucleotides are identified on each cDNA for unique representation in the genome of the organism from which the library is derived. For example, if the library is derived from mammalian mRNA then a randomly selected sequence 14-15 nucleotides long is expected to have unique representation among the 2-3 thousand megabases of the typical mammalian genome. Of course identification of far fewer nucleotides would be sufficient for unique representation in a library derived from bacteria, or other lower organisms. Preferably, at least 20-30 nucleotides are identified to ensure unique representation and to permit construction of a suitable primer as described below. The tabulated sequences may then be compared to known sequences to identify unique cDNAs.

Unique cDNAs are then isolated by conventional techniques, e.g. constructing a probe from the PCR amplicon produced with primers directed to the prime site and the portion of the cDNA whose sequence was determined. The probe may then be used to identify the cDNA in a library using a conventional screening protocol.

The above method for identifying new cDNAs may also be used to fingerprint mRNA populations, either in isolated measurements or in the context of a dynamically changing population. Partial sequence information is obtained simultaneously from a large sample, e.g. ten to a hundred thousand, or more, of cDNAs attached to separate microparticles as described in the above method.

Example 1**Construction of a Tag Library**

An exemplary tag library is constructed as follows to form the chemically
 5 synthesized 9-word tags of nucleotides A, G, and T defined by the formula:



where "[⁴(A,G,T)₉]" indicates a tag mixture where each tag consists of nine 4-mer
 10 words of A, G, and T; and "p" indicate a 5' phosphate. This mixture is ligated to the
 following right and left primer binding regions (SEQ ID NO: 4 and SEQ ID NO 5):

15 5' - AGTGGCTGGGCATCGGACCG TCACCGACCCGTAGCCp LEFT	5' - GGGGCCCAGTCAGCGTCGAT GGGTCAGTCGCAGCTA RIGHT
--	--

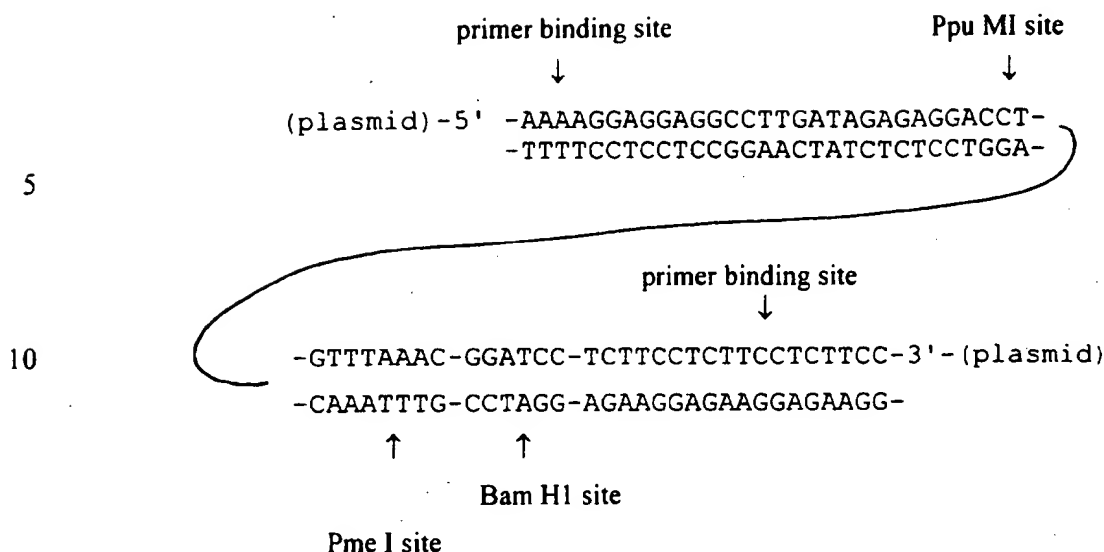
The right and left primer binding regions are ligated to the above tag mixture, after
 which the single stranded portion of the ligated structure is filled with DNA
 20 polymerase then mixed with the right and left primers indicated below and amplified
 to give a tag library (SEQ ID NO: 6).

25 Left Primer 5' - AGTGGCTGGGCATCGGACCG 5' - AGTGGCTGGGCATCGGACCG- TCACCGACCCGTAGCCTGGC-	[⁴ (A, G, T) ₉] - GGGGCCCAGTCAGCGTCGAT [⁴ (A, G, T) ₉] - CCCC GGGTCAGTCGCAGCTA CCCC GGGTCAGTCGCAGCTA - 5'
--	---

Right Primer

35 The underlined portion of the left primer binding region indicates a Rsr II recognition
 site. The left-most underlined region of the right primer binding region indicates
 recognition sites for Bsp 120I, Apa I, and Eco O 109I, and a cleavage site for Hga I.
 The right-most underlined region of the right primer binding region indicates the
 recognition site for Hga I. Optionally, the right or left primers may be synthesized
 40 with a biotin attached (using conventional reagents, e.g. available from Clontech
 Laboratories, Palo Alto, CA) to facilitate purification after amplification and/or
 cleavage.

NOT FURNISHED UPON FILING



The plasmid is cleaved with Ppu MI and Pme I (to give a Rsr II-compatible end and a flush end so that the insert is oriented) and then methylated with DAM methylase. The tag-containing construct is cleaved with Rsr II and then ligated to the open plasmid, after which the conjugate is cleaved with Mbo I and Bam HI to permit ligation and closing of the plasmid. The plasmid is then amplified and isolated and used in accordance with the invention.

Example 3

Changes in Gene Expression Profiles in Liver Tissue of Rats

Exposed to Various Xenobiotic Agents

In this experiment, to test the capability of the method of the invention to detect genes induced as a result of exposure to xenobiotic compounds, the gene expression profile of rat liver tissue is examined following administration of several compounds known to induce the expression of cytochrome P-450 isoenzymes. The results obtained from the method of the invention are compared to results obtained from reverse transcriptase PCR measurements and immunochemical measurements of the cytochrome P-450 isoenzymes. Protocols and materials for the latter assays are described in Morris et al, Biochemical Pharmacology, 52: 781-792 (1996).

Male Sprague-Dawley rats between the ages of 6 and 8 weeks and weighing 200-300 g are used, and food and water are available to the animals *ad lib*. Test compounds are phenobarbital (PB), metyrapone (MET), dexamethasone (DEX), clofibrate (CLO), corn oil (CO), and β -naphthoflavone (BNF), and are available from Sigma Chemical Co. (St. Louis, MO). Antibodies against specific P-450 enzymes are available from the following sources: rabbit anti-rat CYP3A1 from Human Biologics, Inc. (Phoenix, AZ); goat anti-rat CYP4A1 from Daiichi Pure Chemicals Co. (Tokyo,

Japan); monoclonal mouse anti-rat CYP1A1, monoclonal mouse anti-rat CYP2C11, goat anti-rat CYP2E1, and monoclonal mouse anti-rat CYP2B1 from Oxford Biochemical Research, Inc. (Oxford, MI). Secondary antibodies (goat anti-rabbit IgG, rabbit anti-goat IgG and goat anti-mouse IgG) are available from Jackson
5 ImmunoResearch Laboratories (West Grove, PA).

Animals are administered either PB (100 mg/kg), BNF (100 mg/kg), MET (100 mg/kg), DEX (100 mg/kg), or CLO (250 mg/kg) for 4 consecutive days via intraperitoneal injection following a dosing regimen similar to that described by Wang et al, Arch. Biochem. Biophys. 290: 355-361 (1991). Animals treated with
10 H₂O and CO are used as controls. Two hours following the last injection (day 4), animals are killed, and the livers are removed. Livers are immediately frozen and stored at -70°C.

Total RNA is prepared from frozen liver tissue using a modification of the method described by Xie et al, Biotechniques, 11: 326-327 (1991). Approximately
15 100-200 mg of liver tissue is homogenized in the RNA extraction buffer described by Xie et al to isolate total RNA. The resulting RNA is reconstituted in diethylpyrocarbonate-treated water, quantified spectrophotometrically at 260 nm, and adjusted to a concentration of 100 µg/ml. Total RNA is stored in
- diethylpyrocarbonate-treated water for up to 1 year at -70°C without any apparent
20 degradation. RT-PCR and sequencing are performed on samples from these preparations.

For sequencing, samples of RNA corresponding to about 0.5 µg of poly(A)⁺ RNA are used to construct libraries of tag-cDNA conjugates following the protocol described in the section entitled "Attaching Tags to Polynucleotides for Sorting onto
25 Solid Phase Supports," with the following exception: the tag repertoire is constructed from six 4-nucleotide words from Table II. Thus, the complexity of the repertoire is 8⁶ or about 2.6 x 10⁵. For each tag-cDNA conjugate library constructed, ten samples of about ten thousand clones are taken for amplification and sorting. Each of the amplified samples is separately applied to a fixed monolayer of about 10⁶ 10 µm
30 diameter GMA beads containing tag complements. That is, the "sample" of tag complements in the GMA bead population on each monolayer is about four fold the total size of the repertoire, thus ensuring there is a high probability that each of the sampled tag-cDNA conjugates will find its tag complement on the monolayer. After the oligonucleotide tags of the amplified samples are rendered single stranded as
35 described above, the tag-cDNA conjugates of the samples are separately applied to the monolayers under conditions that permit specific hybridization only between oligonucleotide tags and tag complements forming perfectly matched duplexes. Concentrations of the amplified samples and hybridization times are selected to

permit the loading of about 5×10^4 to 2×10^5 tag-cDNA conjugates on each bead where perfect matches occur. After ligation, 9-12 nucleotide portions of the attached cDNAs are determined in parallel by the single base sequencing technique described by Brenner in International patent application PCT/US95/03678. Frequency
5 distributions for the gene expression profiles are assembled from the sequence information obtained from each of the ten samples.

RT-PCRs of selected mRNAs corresponding to cytochrome P-450 genes and the constitutively expressed cyclophilin gene are carried out as described in Morris et al (cited above). Briefly, a 20 μ L reaction mixture is prepared containing 1x reverse
10 transcriptase buffer (Gibco BRL), 10 nM dithiothreitol, 0.5 nM dNTPs, 2.5 μ M oligo d(T)₁₅ primer, 40 units RNasin (Promega, Madison, WI), 200 units RNase H-reverse transcriptase (Gibco BRL), and 400 ng of total RNA (in diethylpyrocarbonate-treated water). The reaction is incubated for 1 hour at 37°C followed by inactivation of the enzyme at 95°C for 5 min. The resulting cDNA is stored at -20°C until used. For
15 PCR amplification of cDNA, a 10 μ L reaction mixture is prepared containing 10x polymerase reaction buffer, 2 mM MgCl₂, 1 unit Taq DNA polymerase (Perkin-Elmer, Norwalk, CT), 20 ng cDNA, and 200 nM concentration of the 5' and 3' specific PCR primers of the sequences described in Morris et al (cited above). PCRs
-are carried out in a Perkin-Elmer 9600 thermal cycler for 23 cycles using melting,
20 annealing, and extension conditions of 94°C for 30 sec., 56°C for 1 min., and 72°C for 1 min., respectively. Amplified cDNA products are separated by PAGE using 5% native gels. Bands are detected by staining with ethidium bromide.

Western blots of the liver proteins are carried out using standard protocols after separation by SDS-PAGE. Briefly, proteins are separated on 10% SDS-PAGE
25 gels under reducing conditions and immunoblotted for detection of P-450 isoenzymes using a modification of the methods described in Harris et al, Proc. Natl. Acad. Sci., 88: 1407-1410 (1991). Protein are loaded at 50 μ g/lane and resolved under constant current (250 V) for approximately 4 hours at 20°C. Proteins are transferred to
nitrocellulose membranes (Bio-Rad, Hercules, CA) in 15 mM Tris buffer containing
30 120 mM glycine and 20% (v/v) methanol. The nitrocellulose membranes are blocked with 2.5% BSA and immunoblotted for P-450 isoenzymes using primary monoclonal and polyclonal antibodies and secondary alkaline phosphatase conjugated anti-IgG. Immunoblots are developed with the Bio-Rad alkaline phosphatase substrate kit.

The three types of measurements of P-450 isoenzyme induction showed
35 substantial agreement.

APPENDIX Ia
Exemplary computer program for generating
minimally cross hybridizing sets
(single stranded tag/single stranded tag complement)

```

Program minxh
C
C
C
      integer*2 sub1(6),mset1(1000,6),mset2(1000,6)
      dimension nbase(6)
C
C
      write(*,*)'ENTER SUBUNIT LENGTH'
      read(*,100)nsup
100    format(i1)
      open(1,file='sub4.dat',form='formatted',status='new')
C
C
      nset=0
      do 7000 m1=1,3
        do 7000 m2=1,3
          do 7000 m3=1,3
            do 7000 m4=1,3
              sub1(1)=m1
              sub1(2)=m2
              sub1(3)=m3
              sub1(4)=m4
C
C
      ndiff=3
C
C
C      Generate set of subunits differing from
C      sub1 by at least ndiff nucleotides.
C      Save in mset1.
C
C
      jj=1
      do 900 j=1,nsup
900    mset1(1,j)=sub1(j)
C
C
      do 1000 k1=1,3
        do 1000 k2=1,3
          do 1000 k3=1,3
            do 1000 k4=1,3
C
C
              nbase(1)=k1
              nbase(2)=k2
              nbase(3)=k3
              nbase(4)=k4

```

```

c
      n=0
      do 1200 j=1,nsub
        if(subl(j).eq.1 .and. nbase(j).ne.1 .or.
1         subl(j).eq.2 .and. nbase(j).ne.2 .or.
3         subf(j).eq.3 .and. nbase(j).ne.3) then
          n=n+1
          endif
1200      continue
c
c
      if(n.ge.ndiff) then
c
c
c          If number of mismatches
c          is greater than or equal
c          to ndiff then record
c          subunit in matrix mset
c
c
c          jj=jj+1
c          do 1100 i=1,nsub
1100      mset1(jj,i)=nbase(i)
c          endif
c
c
c          continue
1000
c
c
c          do 1325 j2=1,nsub
c          mset2(1,j2)=mset1(1,j2)
1325      mset2(2,j2)=mset1(2,j2)
c
c
c          Compare subunit 2 from
c          mset1 with each successive
c          subunit in mset1, i.e. 3,
c          4,5, ... etc. Save those
c          with mismatches .ge. ndiff
c          in matrix mset2 starting at
c          position 2.
c          Next transfer contents
c          of mset2 into mset1 and
c          start
c          comparisons again this time
c          starting with subunit 3.
c          Continue until all subunits
c          undergo the comparisons.
c
c
c          npass=0
c
c
c          continue
1700      kk=npass+2
c          npass=npass+1
c

```

```

c
do 1500 m=npass+2,jj
  n=0
  do 1600 j=1,nsup
    if(mset1(npass+1,j).eq.1.and.mset1(m,j).ne.1.or.
2      mset1(npass+1,j).eq.2.and.mset1(m,j).ne.2.or.
2      mset1(npass+1,j).eq.3.and.mset1(m,j).ne.3) then
      n=n+1
    endif
1600    continue
    if(n.ge.ndiff) then
      kk=kk+1
      do 1625 i=1,nsup
1625        mset2(kk,i)=mset1(m,i)
      endif
1500    continue
  c
  c
  c
  c
  c
  c
  c
  c
  c
  do 2000 k=1,kk
    do 2000 m=1,nsup
2000      mset1(k,m)=mset2(k,m)
    if(kk.lt.jj) then
      jj=kk
      goto 1700
    endif
  c
  c
  nset=nset+1
  write(1,7009)
7009  format(/)
  do 7008 k=1,kk
7008    write(1,7010) (mset1(k,m),m=1,nsup)
7010  format(4i1)
  write(*,*)
  write(*,120) kk,nset
120  format(1x,'Subunits in set=',i5,2x,'Set No=',i5)
7000  continue
  close(1)
  c
  c
end
c
c
*****
*****

```


APPENDIX Ib
Exemplary computer program for generating
minimally cross hybridizing sets
(single stranded tag/single stranded tag complement)

```

Program tagN
C
C
C      Program tagN generates minimally cross-hybridizing
C      sets of subunits given i) N--subunit length, and ii)
C      an initial subunit sequence. tagN assumes that only
C      3 of the four natural nucleotides are used in the tags.
C
C
C      character*1 sub1(20)
C      integer*2 mset(10000,20), nbase(20)
C
C
C      write(*,*)'ENTER SUBUNIT LENGTH'
C      read(*,100)nsub
100    format(i2)
C
C
C      write(*,*)'ENTER SUBUNIT SEQUENCE'
C      read(*,110)(sub1(k),k=1,nsub)
110    format(20a1)
C
C
C      ndiff=10
C
C
C      Let a=1 c=2 g=3 & t=4
C
C
C      do 800 kk=1,nsub
C      if(sub1(kk).eq.'a') then
C        mset(1,kk)=1
C      endif
C      if(sub1(kk).eq.'c') then
C        mset(1,kk)=2
C      endif
C      if(sub1(kk).eq.'g') then
C        mset(1,kk)=3
C      endif
C      if(sub1(kk).eq.'t') then
C        mset(1,kk)=4
C      endif
800    continue
C
C
C      Generate set of subunits differing from
C      sub1 by at least ndiff nucleotides.
C
C
C      jj=1
C
C
C      do 1000 k1=1,3

```

```

do 1000 k2=1,3
  do 1000 k3=1,3
    do 1000 k4=1,3
      do 1000 k5=1,3
        do 1000 k6=1,3
          do 1000 k7=1,3
            do 1000 k8=1,3
              do 1000 k9=1,3
                do 1000 k10=1,3
do 1000 k11=1,3
  do 1000 k12=1,3
    do 1000 k13=1,3
      do 1000 k14=1,3
        do 1000 k15=1,3
          do 1000 k16=1,3
            do 1000 k17=1,3
              do 1000 k18=1,3
                do 1000 k19=1,3
                  do 1000 k20=1,3

c
c

      nbase(1)=k1
      nbase(2)=k2
      nbase(3)=k3
      nbase(4)=k4
      nbase(5)=k5
      nbase(6)=k6
      nbase(7)=k7
      nbase(8)=k8
      nbase(9)=k9
      nbase(10)=k10
      nbase(11)=k11
      nbase(12)=k12
      nbase(13)=k13
      nbase(14)=k14
      nbase(15)=k15
      nbase(16)=k16
      nbase(17)=k17
      nbase(18)=k18
      nbase(19)=k19
      nbase(20)=k20

c
c

do 1250 nn=1,jj

  n=0
  do 1200 j=1,nsup
    if(mset(nn,j).eq.1 .and. nbase(j).ne.1 .or.
1      mset(nn,j).eq.2 .and. nbase(j).ne.2 .or.
2      mset(nn,j).eq.3 .and. nbase(j).ne.3 .or.
3      mset(nn,j).eq.4 .and. nbase(j).ne.4) then
      n=n+1
    endif
1200    continue
c
c

    if(n.lt.ndiff) then
      goto 1000
    endif
1250    continue
c
c

    jj=jj+1
    write(*,130) (nbase(i),i=1,nsup),jj
    do 1100 i=1,nsup

```

```

                                mset(jj,i)=nbase(i)
1100                            continue
C
C
1000    continue
C
C
                                write(*,*)
130      format(10x,20(1x,i1),5x,i5)
                                write(*,*)
                                write(*,120) jj
120      format(1x,'Number of words=',i5)
C
C
                                end
C
C
                                *****
C                                *****
```

APPENDIX 1c
Exemplary computer program for generating
minimally cross hybridizing sets
(double stranded tag/single stranded tag complement)

```

Program 3tagN
C
C
C      Program 3tagN generates minimally cross-hybridizing
C      sets of duplex subunits given i) N--subunit length,
C      and ii) an initial homopurine sequence.
C
C
C      character*1 sub1(20)
C      integer*2 mset(10000,20), nbase(20)
C
C
C      write(*,*) 'ENTER SUBUNIT LENGTH'
C      read(*,100) nsub
100  format(i2)
C
C
C      write(*,*) 'ENTER SUBUNIT SEQUENCE a & g only'
C      read(*,110) (sub1(k), k=1, nsub)
110  format(20a1)
C
C      ndiff=10
C
C      Let a=1 and g=2
C
C      do 800 kk=1, nsub
C      if (sub1(kk).eq.'a') then
C          mset(1, kk)=1
C      endif
C          if (sub1(kk).eq.'g') then
C              mset(1, kk)=2
C          endif
800  continue
C
C      jj=1
C
C      do 1000 k1=1, 3
C      do 1000 k2=1, 3
C      do 1000 k3=1, 3
C      do 1000 k4=1, 3
C      do 1000 k5=1, 3
C      do 1000 k6=1, 3
C      do 1000 k7=1, 3
C      do 1000 k8=1, 3
C      do 1000 k9=1, 3
C      do 1000 k10=1, 3
C      do 1000 k11=1, 3
C      do 1000 k12=1, 3
C      do 1000 k13=1, 3
C      do 1000 k14=1, 3
C      do 1000 k15=1, 3
C      do 1000 k16=1, 3
C      do 1000 k17=1, 3
C      do 1000 k18=1, 3

```

```

do 1000 k19=1,3
do 1000 k20=1,3

c
nbase(1)=k1
nbase(2)=k2
nbase(3)=k3
nbase(4)=k4
nbase(5)=k5
nbase(6)=k6
nbase(7)=k7
nbase(8)=k8
nbase(9)=k9
nbase(10)=k10
nbase(11)=k11
nbase(12)=k12
nbase(13)=k13
nbase(14)=k14
nbase(15)=k15
nbase(16)=k16
nbase(17)=k17
nbase(18)=k18
nbase(19)=k19
nbase(20)=k20

c
do 1250 nn=1,jj

c
n=0
do 1200 j=1,nsup
  if(mset(nn,j).eq.1 .and. nbase(j).ne.1 .or.
1    mset(nn,j).eq.2 .and. nbase(j).ne.2 .or.
2    mset(nn,j).eq.3 .and. nbase(j).ne.3 .or.
3    mset(nn,j).eq.4 .and. nbase(j).ne.4) then
    n=n+1
  endif
1200  continue
c
  if(n.lt.ndiff) then
    goto 1000
  endif
1250  continue
c
  jj=jj+1
  write(*,130) (nbase(i),i=1,nsup),jj
  do 1100 i=1,nsup
    mset(jj,i)=nbase(i)
1100  continue
c
1000  continue
c
  write(*,*)
130  format(10x,20(1x,i1),5x,i5)
  write(*,*)
  write(*,120) jj
120  format(1x,'Number of words=',i5)
c
c
end

```

SEQUENCE LISTING

(1) GENERAL INFORMATION:

(i) APPLICANT: David W. Martin, Jr.

(ii) TITLE OF INVENTION: Measurement of Gene Expression profiles in Toxicity Determination

(iii) NUMBER OF SEQUENCES: 7

(iv) CORRESPONDENCE ADDRESS:

(A) ADDRESSEE: Stephen C. Macevicz, Lynx Therapeutics, Inc.
(B) STREET: 3832 Bay Center Place
(C) CITY: Hayward
(D) STATE: California
(E) COUNTRY: USA
(F) ZIP: 94545

(v) COMPUTER READABLE FORM:

(A) MEDIUM TYPE: 3.5 inch diskette
(B) COMPUTER: IBM compatible
(C) OPERATING SYSTEM: Windows 3.1
(D) SOFTWARE: Microsoft Word 5.1

(vi) CURRENT APPLICATION DATA:

(A) APPLICATION NUMBER:
(B) FILING DATE:
(C) CLASSIFICATION:

(vii) PRIOR APPLICATION DATA:

(A) APPLICATION NUMBER: PCT/US96/09513
(B) FILING DATE: 06-JUN-96

(vii) PRIOR APPLICATION DATA:

(A) APPLICATION NUMBER: PCT/US95/12791
(B) FILING DATE: 12-OCT-95

(viii) ATTORNEY/AGENT INFORMATION:

(A) NAME: Stephen C. Macevicz
(B) REGISTRATION NUMBER: 30,285
(C) REFERENCE/DOCKET NUMBER: 813wo

(ix) TELECOMMUNICATION INFORMATION:

(A) TELEPHONE: (510) 670-9365
(B) TELEFAX: (510) 670-9302

(2) INFORMATION FOR SEQ ID NO: 1:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 11 nucleotides
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 1:

CTAGTCGACC A

11

(2) INFORMATION FOR SEQ ID NO: 2:

- (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 11 nucleotides
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 2:

NRRGATCYNN N

11

(2) INFORMATION FOR SEQ ID NO: 3:

- (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 38 nucleotides
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 3:

GAGGATGCCT TTATGGATCC ACTCGAGATC CCAATCCA

38

(2) INFORMATION FOR SEQ ID NO: 4:

- (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 20 nucleotides
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: double
 (D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 4:

AGTGGCTGGG CATCGGACCG

20

(2) INFORMATION FOR SEQ ID NO: 5:

- (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 20 nucleotides
 (B) TYPE: nucleic acid

(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 5:

GGGGCCCACT CAGCGTCGAT

20

(2) INFORMATION FOR SEQ ID NO: 6:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 20 nucleotides
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 6:

ATCGACGCTG ACTGGGCCCC

16

(2) INFORMATION FOR SEQ ID NO: 7:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 62 nucleotides
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 7:

AAAAGGAGGA GGCCTTGATA GAGAGGACCT GTTTAAACGG ATCCTCTTCC
TCTTCCTCTT CC

50

62

I claim:

1. A method of determining the toxicity of a compound, the method comprising the steps of:
 - 5 administering the compound to a test organism;
extracting a population of mRNA molecules from each of one or more tissues of the test organism;
forming a separate population of cDNA molecules from each population of mRNA molecules from the one or more tissues such that each cDNA molecule of a
10 separate population has an oligonucleotide tag attached, the oligonucleotide tags being selected from the same minimally cross-hybridizing set;
separately sampling each population of cDNA molecules such that substantially all different cDNA molecules within a separate population have different oligonucleotide tags attached;
15 sorting the cDNA molecules of each separate population by specifically hybridizing the oligonucleotide tags with their respective complements, the respective complements being attached as uniform populations of substantially identical complements in spatially discrete regions on one or more solid phase supports;
determining the nucleotide sequence of a portion of each of the sorted cDNA
20 molecules of each separate population to form a frequency distribution of expressed genes for each of the one or more tissues; and
correlating the frequency distribution of expressed genes in each of the one or more tissues with the toxicity of the compound.
- 25 2. The method of claim 1 wherein said oligonucleotide tag and said complement of said oligonucleotide tag are single stranded.
3. The method of claim 2 wherein said oligonucleotide tag consists of a plurality of subunits, each subunit consisting of an oligonucleotide of 3 to 9 nucleotides in
30 length and each subunit being selected from the same minimally cross-hybridizing set.
4. The method of claim 3 wherein said one or more solid phase supports are microparticles and wherein said step of sorting said cDNA molecules onto the microparticles produces a subpopulation of loaded microparticles and a subpopulation
35 of unloaded microparticles.
5. The method of claim 4 further including a step of separating said loaded microparticles from said unloaded microparticles.

6. The method of claim 5 further including a step of repeating said steps of sampling, sorting, and separating until a number of said loaded microparticles is accumulated is at least 10,000.
- 5 7. The method of claim 6 wherein said number of loaded microparticles is at least 100,000.
8. The method of claim 7 wherein said number of loaded microparticles is at
10 least 500,000.
9. The method of claim 5 further including a step of repeating said steps of sampling, sorting, and separating until a number of said loaded microparticles is accumulated is sufficient to estimate the relative abundance of a cDNA molecule
15 present in said population at a frequency within the range of from 0.1% to 5% with a 95% confidence limit no larger than 0.1% of said population.
10. The method of claim 4 wherein said test organism is a mammalian tissue
- culture.
20
11. The method of claim 10 wherein said mammalian tissue culture comprises hepatocytes.
12. The method of claim 4 wherein said test organism is an animal selected from
25 the group consisting of rats, mice, hamsters, guinea pigs, rabbits, cats, dogs, pigs, and monkeys.
13. The method of claim 12 wherein said one or more tissues are selected from the group consisting of liver, kidney, brain, cardiovascular, thyroid, spleen, adrenal, large
30 intestine, small intestine, pancreas urinary bladder, stomach, ovary, testes, and mesenteric lymph nodes.
14. A method of identifying genes which are differentially expressed in a selected
35 tissue of a test animal after treatment with a compound, the method comprising the steps of:
administering the compound to a test animal;

extracting a population of mRNA molecules from the selected tissue of the test animal;

forming a population of cDNA molecules from the population of mRNA molecules such that each cDNA molecule has an oligonucleotide tag attached, the oligonucleotide tags being selected from the same minimally cross-hybridizing set;

sampling the population of cDNA molecules such that substantially all different cDNA molecules have different oligonucleotide tags attached;

sorting the cDNA molecules by specifically hybridizing the oligonucleotide tags with their respective complements, the respective complements being attached as uniform populations of substantially identical complements in spatially discrete regions on one or more solid phase supports;

determining the nucleotide sequence of a portion of each of the sorted cDNA molecules to form a frequency distribution of expressed genes; and

identifying genes expressed in response to administering the compound by comparing the frequency distribution of expressed genes of the selected tissue of the test animal with a frequency distribution of expressed genes of the selected tissue of a control animal.

15. The method of claim 14 wherein said oligonucleotide tag and said complement of said oligonucleotide tag are single stranded.

16. The method of claim 15 wherein said oligonucleotide tag consists of a plurality of subunits, each subunit consisting of an oligonucleotide of 3 to 9 nucleotides in length and each subunit being selected from the same minimally cross-hybridizing set.

17. The method of claim 16 wherein said one or more solid phase supports are microparticles and wherein said step of sorting said cDNA molecules onto the microparticles produces a subpopulation of loaded microparticles and a subpopulation of unloaded microparticles.

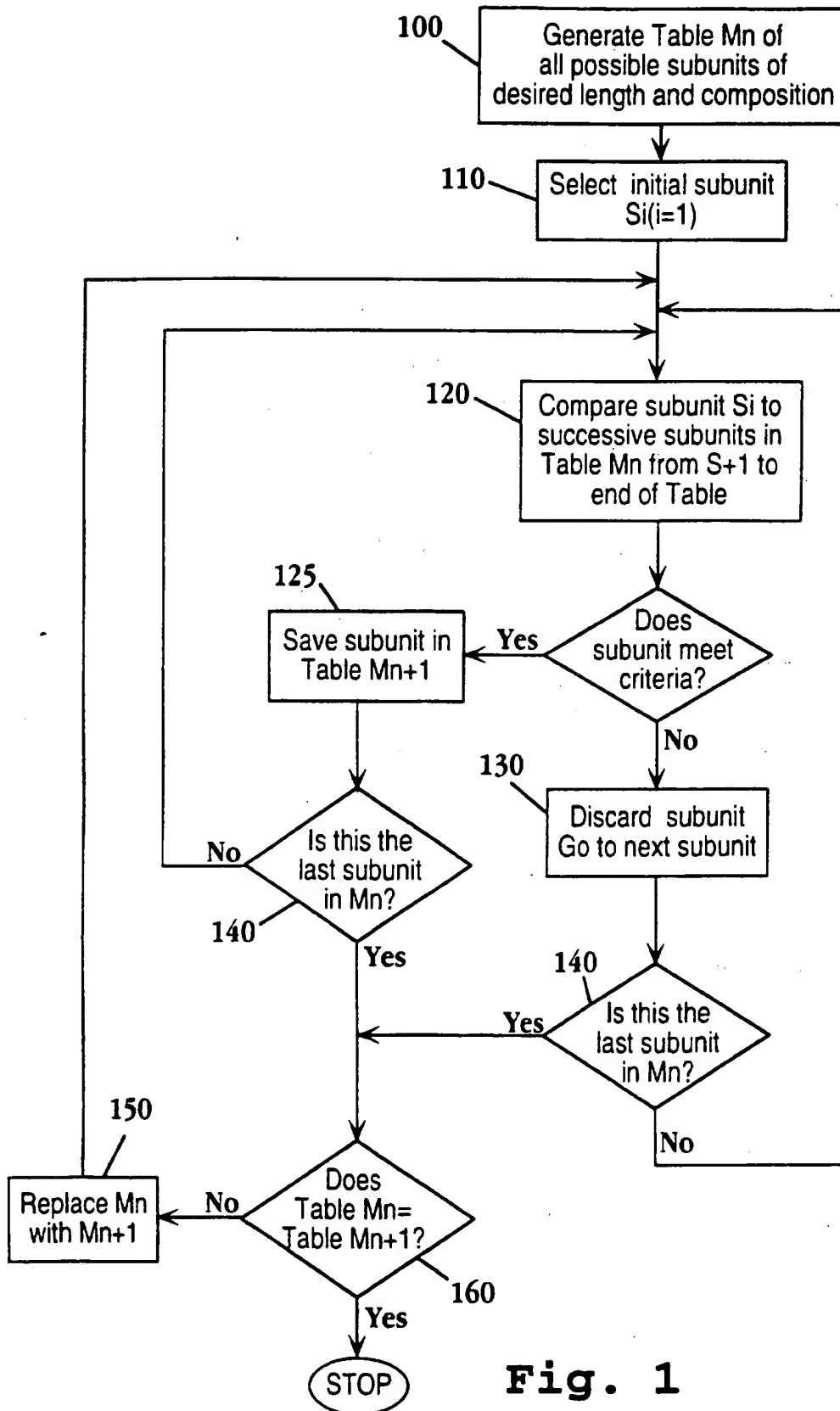
18. The method of claim 17 further including a step of separating said loaded microparticles from said unloaded microparticles.

19. The method of claim 18 further including a step of repeating said steps of sampling, sorting, and separating until a number of said loaded microparticles is accumulated is at least 10,000.

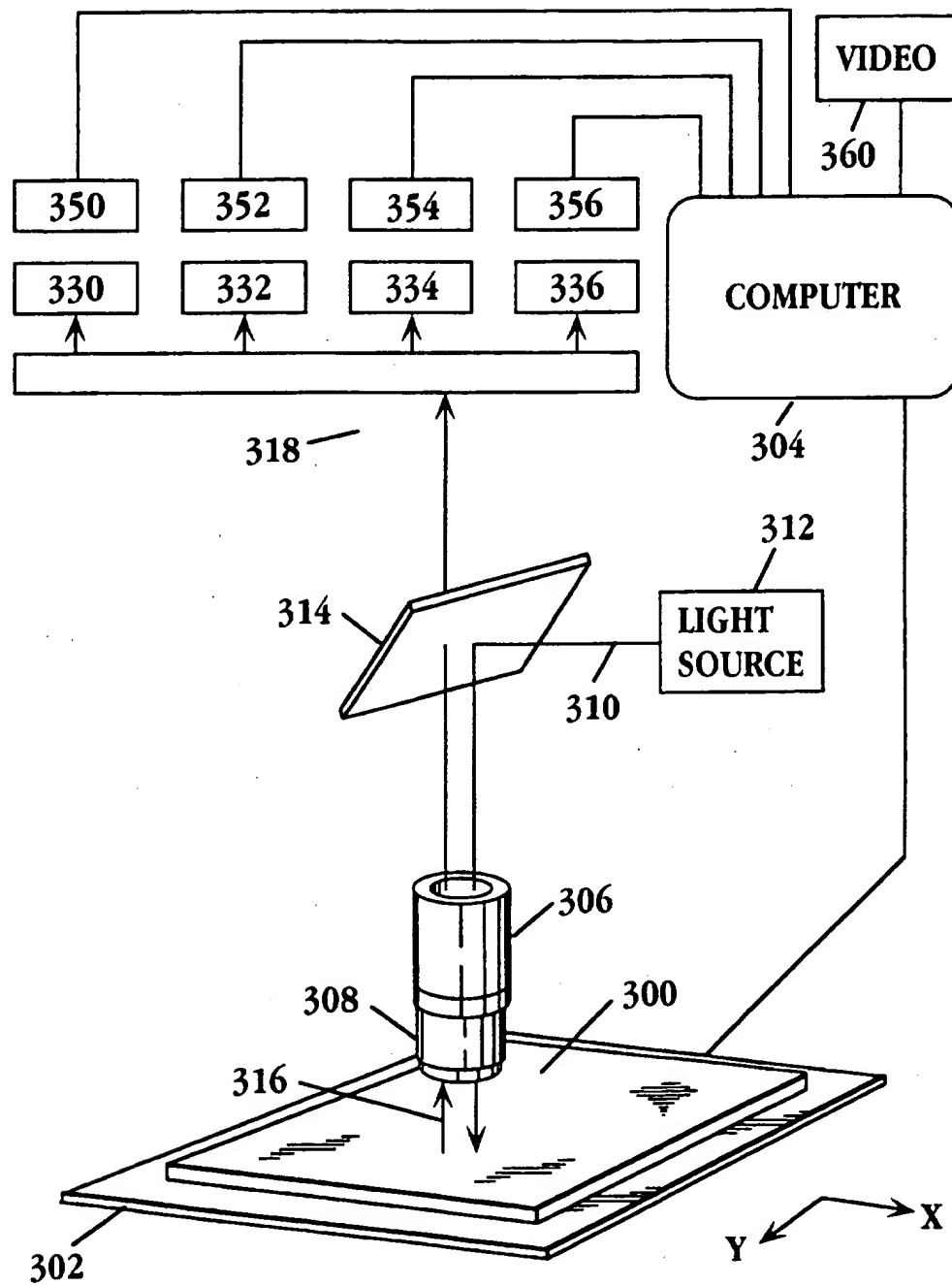
20. The method of claim 19 wherein said number of loaded microparticles is at least 100,000.
21. The method of claim 20 wherein said number of loaded microparticles is at least 500,000.
22. The method of claim 18 further including a step of repeating said steps of sampling, sorting, and separating until a number of said loaded microparticles is accumulated is sufficient to estimate the relative abundance of a cDNA molecule present in said population at a frequency within the range of from 0.1% to 5% with a 95% confidence limit no larger than 0.1% of said population.
23. The method of claim 17 wherein said test animal is selected from the group consisting of rats, mice, hamsters, guinea pigs, rabbits, cats, dogs, pigs, and monkeys.
24. The method of claim 23 wherein said selected tissue is selected from the group consisting of liver, kidney, brain, cardiovascular, thyroid, spleen, adrenal, large intestine, small intestine, pancreas urinary bladder, stomach, ovary, testes, and mesenteric lymph nodes.
25. A use of the technique of massively parallel signature sequencing to determine the toxicity of a compound in a test organism, the use comprising the steps of:
administering the compound to a test organism;
extracting a population of mRNA molecules from each of one or more tissues of the test organism and forming a population of cDNA molecules for each of the one or more tissues;
determining the nucleotide sequence of a portion of each of the cDNA molecules of each separate population using massively parallel signature sequencing to form a frequency distribution of expressed genes for each of the one or more tissues; and
correlating the frequency distribution of expressed genes in each of the one or more tissues with the toxicity of the compound.
26. The use of claim 25 wherein said test organism is a mammalian tissue culture.
27. The use of claim 26 wherein said mammalian tissue culture comprises hepatocytes.

28. The use of claim 25 wherein said test organism is an animal selected from the group consisting of rats, mice, hamsters, guinea pigs, rabbits, cats, dogs, pigs, and monkeys.
- 5 29. The use of claim 28 wherein said one or more tissues are selected from the group consisting of liver, kidney, brain, cardiovascular, thyroid, spleen, adrenal, large intestine, small intestine, pancreas urinary bladder, stomach, ovary, testes, and mesenteric lymph nodes.
- 10 30. A use of the technique of massively parallel signature sequencing to identify genes which are differentially expressed in a test organism after treatment with a compound and which are correlated with toxicity of the compound, the use comprising the steps of:
- administering the compound to the test organism;
 - 15 extracting a population of mRNA molecules from a selected tissue of the test organism and forming a population of cDNA molecules;
 - determining the nucleotide sequence of a portion of each of the cDNA molecules using massively parallel signature sequencing to form a frequency
 - distribution of expressed genes;
 - 20 identifying genes expressed in response to administering the compound by comparing the frequency distribution of expressed genes of the selected tissue of the test organism with a frequency distribution of expressed genes of the selected tissue of a control organism; and
 - determining whether the genes expressed in response to administering the
 - 25 compound are correlated with toxicity of the compound in the test organism.

1/2

**Fig. 1**

2/2

**Fig. 2**

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US96/16342**A. CLASSIFICATION OF SUBJECT MATTER**

IPC(6) : C12Q 1/68; C07H 21/04

US CL : 435/6; 536/24.3

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 435/6; 536/24.3

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

APS, MEDLINE, BIOSIS, CAPLUS, SCISEARCH

search terms: Martin, David W., toxic?, differential?, express?, cDNA, mRNA, RNA, gene#, hybrid?,

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CHETVERIN et al. Oligonucleotide arrays: New concepts and possibilities. Bio/Technology. 12 November 1994, Vol. 12, pages 1093-1099, especially pages 1095-1096.	1-30
A	BRENNER et al. Encoded combinatorial chemistry. Proceedings of the National Academy of Sciences USA. June 1992, Vol. 89, pages 5381-5383.	1-30
A	MATSUBARA et al. cDNA analyses in the human genome project. Gene. 15 December 1993, Vol. 135, No. 1-2, pages 265-274.	1-30

☒ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:	*T	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
A document defining the general state of the art which is not considered to be of particular relevance	*X*	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
E earlier document published on or after the international filing date	*Y*	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
L document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	*G*	document member of the same patent family
O document referring to an oral disclosure, use, exhibition or other means		
P document published prior to the international filing date but later than the priority date claimed		

Date of the actual completion of the international search

27 JANUARY 1997

Date of mailing of the international search report

19 FEB 1997

Name and mailing address of the ISA/US
Commissioner of Patents and Trademarks
Box PCT
Washington, D.C. 20231

Facsimile No. (703) 305-3230

Authorized officer

SCOTT D. PRIEBE

Telephone No. (703) 308-0196

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US96/16342

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	WO 95/21944 A1 (SMITHKLINE BEECHAM CORPORATION) 17 August 1995, page 4, lines 1-4, page 5, lines 31-37, page 17, lines 15-27, page 18, lines 30-35, page 20, line 23 to page 21, line 4.	1-30

FOCUS - 17 of 19 DOCUMENTS

Copyright 1997 PR Newswire Association, Inc.
PR Newswire

August 11, 1997, Monday

SECTION: Financial News

DISTRIBUTION: TO BUSINESS AND MEDICAL EDITORS

LENGTH: 478 words

HEADLINE: Eli Lilly & Co. and Acacia Biosciences Enter Into Research Collaboration;
First Corporate Agreement for Acacia's Genome Reporter Matrix(TM)

DATELINE: RICHMOND, Calif., Aug. 11

BODY:

Acacia Biosciences and Eli Lilly and Company (Lilly) announced today the signing of a joint research collaboration to utilize Acacia's Genome Reporter Matrix(TM) (GRM) to aid in the selection and optimization of lead compounds. Under the collaboration, Acacia will provide chemical and biological profiles on a class of Lilly's compounds for an undisclosed fee.

Acacia's GRM is an assay-based computer modeling system that uses yeast as a miniature ecosystem. The GRM can profile the extent, nature and quantity of any changes in gene expression. Because of the similarities between the yeast and human genome, the system serves as an excellent surrogate for the human body, mimicking the effects induced by a biologically active molecule.

"Using yeast as a model organism for lead optimization makes a lot of sense given the high degree of homology with human metabolic pathways," said William Current of Lilly Research Laboratories. "Acacia's innovative GRM has the potential to provide enormous insight into the therapeutic impact of our compounds and make the drug discovery process more rational. It should substantially accelerate the development process."

"This first agreement with a major pharmaceutical company is an important milestone in the development of Acacia," said Bruce Cohen, President and CEO of Acacia. "The deal is in line with our strategy of establishing alliances that will allow our collaborators to use genomic profiles to identify and optimize compounds within their existing portfolios. In the long run, this technology can be used to characterize large scale combinatorial libraries, predict side effects prior to clinical trials and resurrect drugs that have failed during clinical trials."

The GRM incorporates two critical elements: chemical response profiles and genetic response profiles. The chemical response profiles measure the change in gene expression caused by potential therapeutics and then rank genes with altered expressions by degree of response. The genetic response profiles measure changes in gene expression caused by mutations in the genes encoding potential targets of pharmaceuticals; these genetic response profiles represent gold standards in drug discovery by defining the response profile expected for drugs with perfect selectivity and specificity. By comparing the two profiles, one can analyze a potential drug candidate's ability to mimic the action of a 'perfect' drug.

Acacia Biosciences is a functional genomics company developing proprietary technologies to enhance the speed and efficacy of drug discovery and development. Acacia's Genome Reporter Matrix capitalizes on the latest advances in genomics and combinatorial chemistry to generate comprehensive profiles of drug candidates' in vivo activity.

SOURCE Acacia Biosciences

CONTACT: Bruce Cohen, President and CEO of Acacia Biosciences, 510-669-2330 ext. 103 or Media: Linda Seaton of Feinstein

LOAD-DATE: August 12, 1997

The Bioreactor Market: Steady Growth Expected

The worldwide market for all bioreactors was valued at \$275 million for 1997, and is expected to be worth \$380 million by 2002.

TI: GENETIC ENGINEERING NEWS
09/25/97

GENETIC ENGINEERING GEN NEWS

Reference 9 of 21
in Response dated 3/31/04
In USSN: 09/831,805

BIOTECHNOLOGY

09/25/97

BIOPROCESS

BIORESEARCH • TECHNOLOGY TRANSFER

Contents

European Biotech Standards Meeting	4
Bioprocess Simulation Packages	13
Trends in Infrastructure Development	14
QC/QA for Small Biotech Firms	16
Advances in Electroporation	19
New Products	21
New GEN Column Drug Discovery Assay Miniaturization	27
Corporate Profiles: Pangea Systems	28
Canada Watch	29
European Roundup	30
Wall Street Outlook	31
Collaborations & Agreements	32
Inside Industry	33
Clinical Trials Update	37
Not Quite	40
People	41
Calendar	41
Marketplace	42

Pharmagene Raises More Capital for Research on Human Tissues

By Sophia Fox

Pharmagene, the Royston, U.K.-based biopharmaceutical company specializing in the use of human biomaterials for drug discovery research, has raised a further £5 million from a group of investors led by 3i and Abacus Nominees. The funding will enable the company to expand both its human biomaterials collection and its capabilities across a range of proprietary platform technologies.

Gordon Baxter, Ph.D., Pharmagene's cofounder and chief operating officer, claimed, "by the end of this year Pharmagene will have access to the largest collection of human RNAs and proteins anywhere in the world, and a range of innovative, yet robust technologies

SEE PHARMAGENE, P. 9

Perkin-Elmer Acquires PerSeptive to Expand Its Capabilities in Gene-Based Drug Discovery

By John Sterling

Perkin-Elmer's (PE; Norwalk, CT) decision last month to acquire PerSeptive Biosystems (Framingham, MA) via a \$360 million stock swap was designed to strengthen PE in terms of broad capabilities in gene-based drug discovery. The company's main goal is to develop new products to improve the integration of genetic and protein research.

"This merger will enhance our position as an effective provider of innovative, integrated platforms enabling our customers to be more efficient and cost-effective in bringing new pharmaceuticals to market," says Tony L. White, PE's chairman, president and CEO. "The combination of our two companies should bolster our presence in the life sciences, [and it is our] belief that we must take bold action now to lead the emerging era of molecular medicine with leading positions in both genetic and protein analysis."

A driving force behind the merger is the vast amount of genetic



Perkin-Elmer acquired PerSeptive Biosystems for \$360 million to obtain new technologies in mass spectrometry, bioseparations and purification for product development projects, spanning the range from genomics to proteomics.

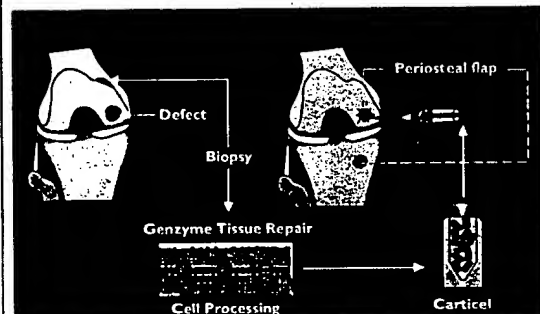
information about human disease that is being accumulated by researchers and biotech companies working in the area of genomics. It is becoming increasingly obvious that these data need to be complemented with technologies for

studying proteins and protein networks—a field known as proteomics (see GEN, September 1, 1997, p. 1).

PE officials, who claim that MALDI-TOF (Matrix Assisted

SEE ACQUISITION, P. 10

FDA OKs Genzyme's Carticel Product for Damage to Knees



Carticel, which was approved for the repair of clinically significant, symptomatic cartilaginous defects of the femoral condyle (medial, lateral or trochlear) caused by acute or repetitive trauma, employs a proprietary process to grow autologous cartilage cells for implantation.

By Naomi Pfeiffer

The FDA has approved a knee-cartilage replacement product made by Genzyme Tissue Repair (Cambridge, MA), a tracking-stock division of Genzyme Corp., for people with trauma-damaged knees.

Carticel™ (autologous cultured chondrocytes) is the first product to be licensed under the FDA's pro-

SEE GENZYME, P. 6

Strategies for Target Validation Streamline Evaluation of Leads

By Vicki Glaser

Acacia Biosciences (Richmond, CA) last month announced its first agreement with a major pharmaceutical company, signing a deal with Eli Lilly (Indianapolis, IN) to use Acacia's Genome Reporter Matrix (GRM) to select and optimize some of Lilly's lead compounds. Acacia's yeast-based system for profiling drug activity is useful for evaluating the therapeutic potential of lead compounds, and it also has a role in the identification and validation of new drug targets.

"We're using the ecosystem of a cell to allow us to deduce the mechanism of action and target for any chemical," explains Bruce Cohen, president and CEO. "We screen for every target in a cell simultaneously...using transcription as a readout

for how a cell is adapting to any perturbation," he says.

The GRM technology consists of two main databases: one is the genetic response profile, showing the effects of mutations in each individual yeast gene and compensatory gene regulatory mechanisms; the other is the chemical response profile, which documents changes in gene expression in response to chemical compounds. Computational analysis and pattern matching between the genetic and chemical profiles yields information on the specificity, potency and side-effects risk of a drug lead.

Targeting Targets

No longer is mapping and sequencing a gene—or the human genome—an end unto itself, but

SEE TARGET, P. 15

Sticky Ends

Avigen received two grants from the NIH & University of California for research on gene therapy for treatment of cancer & HIV infections...MRL Pharmaceutical Services, of Reston, VA, launched the TSN Bug Finder, which is able to locate & retrieve client-specified microorganisms in real-time...Genasia Sidor, Inc. will move its corporate staff from San Diego to Irvine, CA, by end of year...

FDA accepted NDA from Sepracor for levalbuterol HCl inhalation solution...An \$11.7M mezzanine financing has been closed by Activated Cell Therapy, which changed its name to Dendreon Corporation...Astra AB will build major research facility in Waltham, MA, and is also relocating Astra Arcus research facility from Rochester to Boston area...Prolifix Ltd. team used a small peptide to inhibit the E2F protein complex and induced

apoptosis in mammalian tumor cells...Verata Pharmaceuticals, Inc. and Alpha Therapeutic Corp. ended an agreement to develop VX-366 for treatment of inherited hemoglobin disorders...Navicore received Phase I SBIR grant for up to \$100,000 from NIH for development of prototype of its Naviflow technology for high-throughput screening...Covance Inc. will invest \$21 million in expansion and renovation of its facility in Indianapolis, IN.



Target

from page 1

merely a means to an end. The critical next step is to validate the gene and its protein product as a potential drug target. The Human Genome Project continues to produce a treasure chest of expressed sequence tags (ESTs) and a tantalizing array of complete gene sequences.

Companies are applying a variety of functional genomic strategies to link genes to specific diseases and to multigenic phenotypes. Yet the ultimate challenge for pharmaceutical companies is to sift through all the sequence and differential gene expression data to identify the best targets for drug discovery.

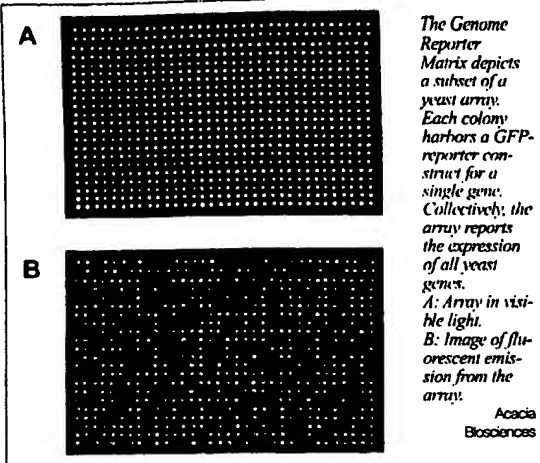
Spinning off technology developed at the University of North Carolina (Chapel Hill), Cytogen Corp. (Princeton, NJ) formed its wholly owned subsidiary AxCell Biosciences earlier this year. The young company is building a protein interaction database, cataloging all the interactions the modular domains of proteins can engage in with a

range of ligands, in order to gain insight into protein function and to select the most critical interaction to target for drug development.

AxCell's cloning-of-ligand-targets (COLT) technology employs "recognition units" from the company's genetic diversity library (ODL) to map functional protein interactions and quantitate their affinity. The company's inter-functional proteomic database (IFP-dbase) elucidates protein interaction networks and structure-activity relationships based on ligand affinity with protein modular domains.

Defining Disease Pathways

Signal Pharmaceuticals, Inc.'s (San Diego, CA) integrated drug target and discovery effort is based on mapping gene-regulating pathways in cells and identifying small molecules that regulate the activation of those genes. In collaboration with academic researchers, the company has identified a large number of regulatory proteins in several mitogen-activated protein (MAP) kinase pathways (including the JNK, FRK and p38



The Genome Reporter Matrix depicts a subset of a yeast array. Each colony harbors a GFP-reporter construct for a single gene. Collectively, the array reports the expression of all yeast genes.

Acacia Biosciences

signaling pathways), which Signal is evaluating for the treatment of autoimmune, inflammatory, cardiovascular and neurologic diseases, and cancer. Other target identification

programs focus on the NF- κ B pathway, estrogen-related genes and central/peripheral nervous system genes.

Regulating cytokine production in immune and inflammatory disorders,

and modifying bone metabolism to treat osteoporosis are the focus of Signal's collaboration with Tanabe Sanyaku (Osaka, Japan). Signal has partnered with Organon/Akzo Nobel (Netherlands) to identify estrogen-responsive genes as targets for treating neurodegenerative and psychiatric diseases, atherosclerosis and ischemia, and with Roche Bioscience (Palo Alto, CA) to develop human peripheral nerve cell lines for the discovery of treatments for pain and incontinence.

Exelixis' (S. San Francisco, CA) strategy for target selection is to define disease pathways and identify regulatory molecules that activate or inhibit those biochemical/genetic pathways. Based on the finding that these pathways are conserved across species, the company is studying the model genetic systems of *Drosophila* and *Caenorhabditis elegans*. Using its PathFinder technology, Exelixis systematically introduces mutations into the genomes of these model organisms, looking for mutations that enhance or suppress the target disease-related gene. These novel genes then become the basis of drug screening assays.

Cadus Pharmaceutical Corp. (Tarrytown, NY) is identifying surrogate ligands to newly discovered orphan G-protein coupled transmembrane receptors of unknown function to determine the suitability of the receptors as drug targets. Inserting the novel receptor in a yeast system yields a ligand that activates the receptor. Access to a surrogate ligand allows the company to screen for receptor antagonists in the yeast system.

"The antagonist plus the surrogate ligand gives you two probes—an on probe and an off probe—which allows you to look at function," explains David Webb, Ph.D., vp of research and chief scientific officer. A surrogate ligand also provides information on which G-protein interacts with the orphan receptor and its associated signaling pathways, further clarifying the role of the receptor as a potential drug target. Cadus' collaboration with SmithKline (Philadelphia) capitalizes on Cadus' ability to determine orphan receptor function, applying the technology to SmithKline's proprietary, newly discovered G-protein receptors.

Cadus' recombinant yeast system can also be used to screen cell and tissue extracts for natural ligands, and the company is accelerating its internal drug-discovery efforts in the areas of cancer, inflammation and allergy. A recent equity investment in Axiom Biotechnologies (San Diego, CA) gave Cadus a license to Axiom's high-throughput pharmacologic screening system for lead optimization and discovery.

As its name implies, **geneNetworks** (Alameda, CA) focuses on identifying gene networks that contribute to multigenic phenotypes and complex disease processes. The integration of mouse and human genetic studies forms the basis of the technology. The Genome Tagged Mice database in development will serve as a library of natural mouse genetic and phenotypic variation. Disease-related genes identified in mice are then evaluated in human family- and population-based studies to confirm their clinical relevance and linkages to pathophysiologic traits.

Blocking Gene Expression

Inactivating a gene known to be expressed in association with a particular disease is one approach to identifying appropriate therapeutic targets. The target validation and discovery program at **Ribozyme Pharmaceuticals, Inc.** (Boulder, CO) applies the company's ribozyme technology to achieve selective inhibition of gene expression in cell culture and in animals.

Correlation of the gene expression inhibition with phenotype can **SEE TARGET, P. 38**

A strong chemical combination to help you grow. And flourish.

Three hundred million dollars and ten years of hard work. That's what it costs to bring your biotechnology-derived therapeutic to the marketplace.

Which means, no room for error.

Which means, in turn, you'd be wise to tap into the combined capabilities of Mallinckrodt and J.T.Baker: dual sources, trusted names for your chemical raw materials.

Two separate GMP-produced brands offering the control of a single quality system and the convenience of a single audit process.

We offer comprehensive product lines including USP salts, bioreagents, high purity solvents and chromatography products in Beaker to Bulk™ packaging for easy scale-up.

Call 1-800-582-2537, or access our website at <http://www.mallinckrodt.com>. For dual chemical sources dedicated to helping you grow. Flourish. Succeed!

MALLINCKRODT



Target

from page 15

suggest the relative importance of the gene in disease pathology. The company's nuclease-resistant ribozymes form the basis of a collaboration with Schering AG (Germany) for drug target validation and the development of ribozyme-based therapeutic agents, and with Chiron Corp. (Emeryville, CA) for target validation.

With several antisense compounds now progressing through clinical trials, the concept of using oligonucleotides to inhibit gene activity is not new. But rather than focusing on therapeutics development, Sequitur, Inc. (Natick, MA) is creating antisense compounds for the purpose of determining gene function and validating drug targets. Clients typically provide the one-year-old company with the sequence (or EST) of a potential gene target and, in return, Sequitur custom designs a series of three to six antisense compounds that yield a three-to-ten-fold inhibition of the target gene in cell culture. The company also provides oligofectins, a series of cationic lipids, to deliver the oligonucleotides to a variety of cultured cells.

"Differential expression information is just for correlation, it doesn't tell function or confirm what would be a good target," says Tod Woolf, Ph.D., director of technology development at Sequitur. Whereas, antisense compounds will inhibit a target, Sequitur offers both phosphorothioate DNA antisense compounds, and its proprietary Next Generation chimeric oligonucleotides, which have a higher hybridization affinity, greater specificity and reduced toxicity, according to the company.

Mining Pathogen Genomes

Companies such as Human Genome Sciences (HGS; Rockville, MD), Incyte (Palo Alto, CA),

ArCell Biosciences scientists say their technology enables the rapid and simple functional identification of the two essential molecular components of protein interaction networks: specific recognition units that bind distinct modular protein domains are identified and isolated using a combination structural/functional approach that uses both peptide phase display Genetic Diversity Libraries (GDL) and bioinformatics, and cloning of Ligand Targets (COLT) technology utilizes recognition units as functional probes to isolate families of interactor proteins.

Millennium Pharmaceuticals Inc. (Cambridge, MA) and Genome Therapeutics (Waltham, MA) are relying on high-speed DNA sequencing, positional cloning and other strategies to identify specific microbial genomic sites that would be good targets for infectious disease therapeutics.

HGS recently completed sequencing of the bacterial pathogen *Streptococcus pneumoniae*, which is the focus of an agreement with Hoffmann-La Roche (Basel, Switzerland). Roche will use the sequence data to develop new anti-infectives against *S. pneumoniae*. HGS and Roche have expanded their collaboration to include a nonexclusive license to access sequence information for the intestinal bacterium *Enterococcus faecalis*.

Incyte Pharmaceuticals has completed one-fold coverage of the *Candida albicans* genome, identifying

60% of the genes of this fungal pathogen. This genome will become part of the company's PathoSeq microbial database. Incyte recently introduced the ZooSeq animal gene sequence and expression database. The database will provide genomic information across various species commonly used in preclinical drug testing, which may help to better define potential drug targets.

Millennium Pharmaceuticals continues to report success in identifying novel drug targets, having recently discovered a novel chemokine called neurotactin and a new class of MAD-related proteins that inhibit transforming growth factor beta (TGF- β) signaling. The company also received U.S. patent coverage for the tub genes, believed to play a role in obesity, and for the gene that encodes the protein melanastin, which appears to suppress metastasis in malignant melanoma.

Pangea

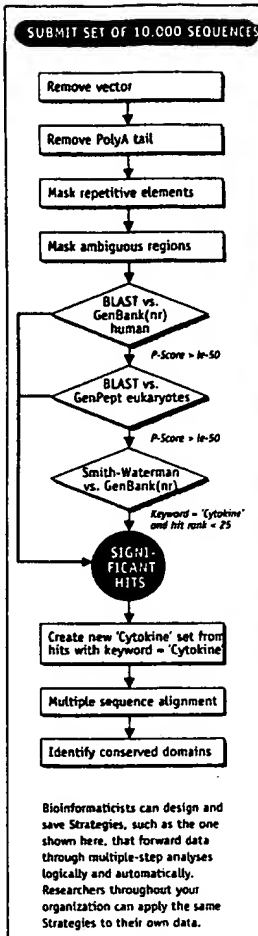
from page 28

Smith, now a computer programmer, is an expert in systems integration, Internet technologies and the application of industrial engineering principles to the drug discovery process. Before co-founding Pangea, he was the manager of software development at Attorney's Briefcase, a legal research software company.

By being "in the trenches" with customers and collaborators, Bellenson and Smith sensed the frustration of pharmaceutical researchers whose incompatible tools have impeded their progress. According to Bellenson, "Most of them are geared toward analyzing one molecule at a time. It's like emptying the ocean with an eye dropper—an incompatible eye dropper at that. A pharmaceutical company may have 30 different drug discovery teams with various approaches. The problem is to manage the process of experimenting with a lot of different approaches, to automate while maintaining flexibility."

GeneWorld 2.1 enables "integration of the entire target discovery and validation process," Bellenson says. The commercial software package coordinates the entire process of sequence-data analysis and can be integrated with other programs and databases, according to Smith, who adds that it handles thousands of sequence results, organizes and automates annotation and seamlessly interacts with growing genome databases. Simple forms and menus enable users to turn raw sequence data into crucial knowledge for drug discovery by applying algorithms to sequences, creating custom analysis strategies and producing useful reports, without the need for writing computer code. GeneWorld 2.1 runs on a variety of platforms and operating systems.

Pairing industrial relational database-management systems with a web-browser interface, Pangea's Operating System of Drug Discovery "is an open-computing framework that allows client/server and Java-enabled web-based technologies to collect, organize and analyze drug discovery information for pharmaceutical companies to simplify and accelerate drug discovery. The technology unites automated genomics database analysis for drug target site selection, chemical information database analysis and large-scale combinatorial chemistry project management and high-throughput screening project management for drug lead efficacy analysis. Pangea officials maintain that these integrated elements provide a unified environment for chemists, biologists and others involved in the drug discovery process to work together with



commercial and public domain software.

Pangea's Operating System of Drug Discovery can accommodate Sybase, Oracle or Informix relational database-management systems and any version of UNIX. It absorbs new data formats, databases, algorithms and analysis paradigms into the automated workflow without software modifications. Netscape Navigator provides a friendly user interface from PC, Macintosh, and UNIX workstations.

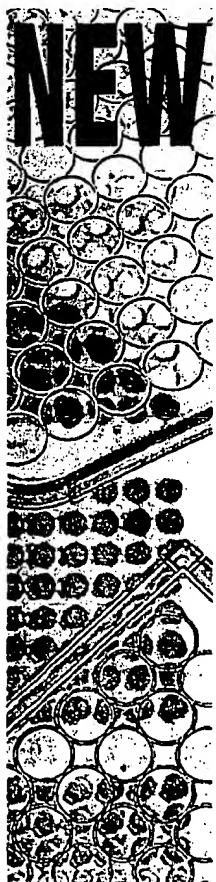
In the near term, Pangea plans to complete its bioinformatics core with two more programs. Gene Foundry, a sample tracking and workflow sequence package for DNA sequence and fragment information, will also offer interaction with robots, reagent tracking and troubleshooting. Gene Thesaurus, the other package is a "warehouse of bioinformatics data," says Bellenson.

breast cancer, breast cancer with liver metastases, glioblastoma, malignant ascites due to gastrointestinal cancer and ovarian cancer.

Copies of the GTAC third annual report are available from the GTAC Secretariat, Wellington House, 133-155 Waterloo Road, London SE1 8UG, U.K.

Coated Lenses Prevent PCO

Scientists in the U.K. say it may be possible to prevent posterior capsule opacification (PCO), a common complication following cataract surgery, by using the implanted polymethylmethacrylate (PMMA) intraocular lens as a drug delivery system. PCO occurs in 30-50% of cataract surgery patients as a result of stimulated cell growth within the remaining capsular bag. The condition causes a decline in visual acuity and requires expensive laser treatment, thus negating the routine use of cataract surgery in underdeveloped countries, explains G. Duncan, at the



NEW HIGH SPECIFIC ACTIVITY MICROBIAL ALKALINE PHOSPHATASE from Biocatalysts

Biocatalysts Limited, the British speciality enzyme company, has developed a completely new type of alkaline phosphatase with many advantages over the types most commonly used.

It is of microbial origin with a high specific activity (unlike that from *E. coli*) and with higher temperature and storage stability compared to that from calf intestine.

This is the first of several new generation diagnostic enzymes being developed by Biocatalysts Limited with greatly improved stability.

- Non-animal source, no risk of BSE or animal virus contamination
- Higher temperature stability than calf intestine
- Much higher specific activity than from *E. coli*
- Very high storage stability even in the absence of glycerol

For further details on alkaline phosphatase and our other diagnostic enzymes contact us direct at the address below or within North America contact our US Distributor Kaltron-Pettibone phone: 630 350 1116 or fax: 630-350-1606

Biocatalysts Limited
Treforest Industrial Estate Pontypridd Wales UK CF37 5UD
Tel: +44 (0)1443 843712 Fax: +44 (0)1443 841214
e-mail: Kelly@Biocatalysts.com.

Europe

from page 30

GTAC Chairman, Professor Norman C. Nevin, said 1996 saw "four important developments": an increase in enquiries and submissions made to GTAC; an increase in the complexity of submitted protocols; a continuing shift from gene therapy for single-gene disorders toward strategies aimed at tumour destruction in cancer; and a growth in international sponsorship of U.K. gene therapy trials.

Since 1993, GTAC and its predecessor, the Clothier Committee, have approved 18 U.K. gene therapy clinical trials (13 of which have been carried out), which are listed in the report. The disease areas targeted by these trials include severe combined immunodeficiency (1 trial), cystic fibrosis (6), metastatic melanoma (2), lymphoma (2), neuroblastoma (1), breast cancer (1), Hurler's syndrome (1), cervical cancer (1), glioblastoma

- Fischer-Vize, *Science* 270, 1828 (1995).
35. T. C. James and S. C. Elgin, *Mol. Cell Biol.* 6, 3862 (1986); R. Paro and D. S. Hogness, *Proc. Natl. Acad. Sci. U.S.A.* 88, 263 (1991); B. Tschiersch *et al.*, *EMBO J.* 13, 3822 (1994); M. T. Madireddi *et al.*, *Cell* 87, 75 (1996); D. G. Stokes, K. D. Tartof, R. P. Perry, *Proc. Natl. Acad. Sci. U.S.A.* 93, 7137 (1996).
 36. P. M. Palosaari *et al.*, *J. Biol. Chem.* 266, 10750 (1991); A. Schmitz, K. H. Gartemann, J. Fiedler, E. Grund, R. Eichenlaub, *Appl. Environ. Microbiol.* 58, 4068 (1992); V. Sharma, K. Suvarna, R. Megathanan, M. E. Hudspeth, *J. Bacteriol.* 174, 5057 (1992); M. Kanazawa *et al.*, *Enzyme Protein* 47, 9 (1993); Z. L. Boynton, G. N. Bennet, F. B. Rudolph, *J. Bacteriol.* 178, 3015 (1996).
 37. M. Ho *et al.*, *Cell* 77, 869 (1994).
 38. W. Hendriks *et al.*, *J. Cell Biochem.* 59, 418 (1995).
 39. We thank H. Skaletsky and F. Lewitter for help with

sequence analysis; Lawrence Livermore National Laboratory for the flow-sorted Y cosmid library; and P. Bain, A. Bortvin, A. de la Chapelle, G. Fink, K. Jegalian, T. Kawaguchi, E. Lander, H. Lodish, P. Matsudaira, D. Menke, U. RajBhandary, R. Reijo, S. Rozen, A. Schwartz, C. Sun, and C. Tilford for comments on the manuscript. Supported by NIH.

28 April 1997; accepted 9 September 1997

Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale

Joseph L. DeRisi, Vishwanath R. Iyer, Patrick O. Brown*

DNA microarrays containing virtually every gene of *Saccharomyces cerevisiae* were used to carry out a comprehensive investigation of the temporal program of gene expression accompanying the metabolic shift from fermentation to respiration. The expression profiles observed for genes with known metabolic functions pointed to features of the metabolic reprogramming that occur during the diauxic shift, and the expression patterns of many previously uncharacterized genes provided clues to their possible functions. The same DNA microarrays were also used to identify genes whose expression was affected by deletion of the transcriptional co-repressor *TUP1* or overexpression of the transcriptional activator *YAP1*. These results demonstrate the feasibility and utility of this approach to genomewide exploration of gene expression patterns.

The complete sequences of nearly a dozen microbial genomes are known, and in the next several years we expect to know the complete genome sequences of several metazoans, including the human genome. Defining the role of each gene in these genomes will be a formidable task, and understanding how the genome functions as a whole in the complex natural history of a living organism presents an even greater challenge.

Knowing when and where a gene is expressed often provides a strong clue as to its biological role. Conversely, the pattern of genes expressed in a cell can provide detailed information about its state. Although regulation of protein abundance in a cell is by no means accomplished solely by regulation of mRNA, virtually all differences in cell type or state are correlated with changes in the mRNA levels of many genes. This is fortuitous because the only specific reagent required to measure the abundance of the mRNA for a specific gene is a cDNA sequence. DNA microarrays, consisting of thousands of individual gene sequences printed in a high-density array on a glass microscope slide (1, 2), provide a practical and economical tool for studying gene expression on a very large scale (3–6).

Saccharomyces cerevisiae is an especially

favorable organism in which to conduct a systematic investigation of gene expression. The genes are easy to recognize in the genome sequence, cis regulatory elements are generally compact and close to the transcription units, much is already known about its genetic regulatory mechanisms, and a powerful set of tools is available for its analysis.

A recurring cycle in the natural history of yeast involves a shift from anaerobic (fermentation) to aerobic (respiration) metabolism. Inoculation of yeast into a medium rich in sugar is followed by rapid growth fueled by fermentation, with the production of ethanol. When the fermentable sugar is exhausted, the yeast cells turn to ethanol as a carbon source for aerobic growth. This switch from anaerobic growth to aerobic respiration upon depletion of glucose, referred to as the diauxic shift, is correlated with widespread changes in the expression of genes involved in fundamental cellular processes such as carbon metabolism, protein synthesis, and carbohydrate storage (7). We used DNA microarrays to characterize the changes in gene expression that take place during this process for nearly the entire genome, and to investigate the genetic circuitry that regulates and executes this program.

Yeast open reading frames (ORFs) were amplified by the polymerase chain reaction (PCR), with a commercially available set of primer pairs (8). DNA microarrays, containing approximately 6400 distinct DNA sequences, were printed onto glass slides by

using a simple robotic printing device (9). Cells from an exponentially growing culture of yeast were inoculated into fresh medium and grown at 30°C for 21 hours. After an initial 9 hours of growth, samples were harvested at seven successive 2-hour intervals, and mRNA was isolated (10). Fluorescently labeled cDNA was prepared by reverse transcription in the presence of Cy3(green)- or Cy5(red)-labeled deoxyuridine triphosphate (dUTP) (11) and then hybridized to the microarrays (12). To maximize the reliability with which changes in expression levels could be discerned, we labeled cDNA prepared from cells at each successive time point with Cy5, then mixed it with a Cy3-labeled "reference" cDNA sample prepared from cells harvested at the first interval after inoculation. In this experimental design, the relative fluorescence intensity measured for the Cy3 and Cy5 fluors at each array element provides a reliable measure of the relative abundance of the corresponding mRNA in the two cell populations (Fig. 1). Data from the series of seven samples (Fig. 2), consisting of more than 43,000 expression-ratio measurements, were organized into a database to facilitate efficient exploration and analysis of the results. This database is publicly available on the Internet (13).

During exponential growth in glucose-rich medium, the global pattern of gene expression was remarkably stable. Indeed, when gene expression patterns between the first two cell samples (harvested at a 2-hour interval) were compared, mRNA levels differed by a factor of 2 or more for only 19 genes (0.3%), and the largest of these differences was only 2.7-fold (14). However, as glucose was progressively depleted from the growth media during the course of the experiment, a marked change was seen in the global pattern of gene expression. mRNA levels for approximately 710 genes were induced by a factor of at least 2, and the mRNA levels for approximately 1030 genes declined by a factor of at least 2. Messenger RNA levels for 183 genes increased by a factor of at least 4, and mRNA levels for 203 genes diminished by a factor of at least 4. About half of these differentially expressed genes have no currently recognized function and are not yet named. Indeed, more than 400 of the differentially expressed genes have no apparent homology

Department of Biochemistry, Stanford University School of Medicine, Howard Hughes Medical Institute, Stanford, CA 94305–5428, USA.

* To whom correspondence should be addressed. E-mail: pbrown@crgm.stanford.edu

to any gene whose function is known (15). The responses of these previously uncharacterized genes to the diauxic shift therefore provides the first small clue to their possible roles.

The global view of changes in expression of genes with known functions provides a vivid picture of the way in which the cell adapts to a changing environment. Figure 3 shows a portion of the yeast metabolic pathways involved in carbon and energy metabolism. Mapping the changes we observed in the mRNAs encoding each enzyme onto this framework allowed us to infer the redirection in the flow of metabolites through this system. We observed large inductions of the genes coding for the enzymes aldehyde dehydrogenase (ALD2) and acetyl-coenzyme A (CoA) synthase (ACS1), which function together to convert the products of alcohol dehydrogenase into acetyl-CoA, which in turn is used to fuel the tricarboxylic acid (TCA) cycle and the glyoxylate cycle. The concomitant shutdown of transcription of the genes encoding pyruvate decarboxylase and induction of pyruvate carboxylase rechannels pyruvate away from acetaldehyde, and instead to oxalacetate, where it can serve to supply the TCA cycle and gluconeogenesis. Induction of the pivotal genes *PCK1*, encoding phosphoenolpyruvate carboxykinase, and *FBP1*, encoding fructose 1,6-bisphosphatase, switches the directions of two key irreversible steps in glycolysis, reversing the flow of metabolites along the reversible steps of the glycolytic pathway toward the essential biosynthetic precursor, glucose-6-phosphate. Induction of the genes coding for the trehalose synthase and glycogen synthase complexes promotes channeling of glucose-6-phosphate into these carbohydrate storage pathways.

Just as the changes in expression of genes encoding pivotal enzymes can provide insight into metabolic reprogramming, the behavior of large groups of functionally related genes can provide a broad view of the systematic way in which the yeast cell adapts to a changing environment (Fig. 4). Several classes of genes, such as cytochrome *c*-related genes and those involved in the TCA/glyoxylate cycle and carbohydrate storage, were coordinately induced by glucose exhaustion. In contrast, genes devoted to protein synthesis, including ribosomal proteins, tRNA synthetases, and translation, elongation, and initiation factors, exhibited a coordinated decrease in expression. More than 95% of ribosomal genes showed at least twofold decreases in expression during the diauxic shift (Fig. 4) (13). A noteworthy and illuminating exception was that the

genes encoding mitochondrial ribosomal genes were generally induced rather than repressed after glucose limitation, highlighting the requirement for mitochondrial biogenesis (13). As more is learned about the functions of every gene in the yeast genome, the ability to gain insight into a cell's response to a changing environment through its global gene expression patterns will become increasingly powerful.

Several distinct temporal patterns of expression could be recognized, and sets of genes could be grouped on the basis of the similarities in their expression patterns. The characterized members of each of these groups also shared important similarities in their functions. Moreover, in most cases, common regulatory mechanisms could be inferred for sets of genes with similar expression profiles. For example, seven genes showed a late induction profile, with mRNA levels increasing by more than ninefold at

the last timepoint but less than threefold at the preceding timepoint (Fig. 5B). All of these genes were known to be glucose-repressed, and five of the seven were previously noted to share a common upstream activating sequence (UAS), the carbon source response element (CSRE) (16–20). A search in the promoter regions of the remaining two genes, *ACR1* and *IDP2*, revealed that *ACR1*, a gene essential for *ACS1* activity, also possessed a consensus CSRE motif, but interestingly, *IDP2* did not. A search of the entire yeast genome sequence for the consensus CSRE motif revealed only four additional candidate genes, none of which showed a similar induction.

Examples from additional groups of genes that shared expression profiles are illustrated in Fig. 5, C through F. The sequences upstream of the named genes in Fig. 5C all contain stress response elements (STRE), and with the exception

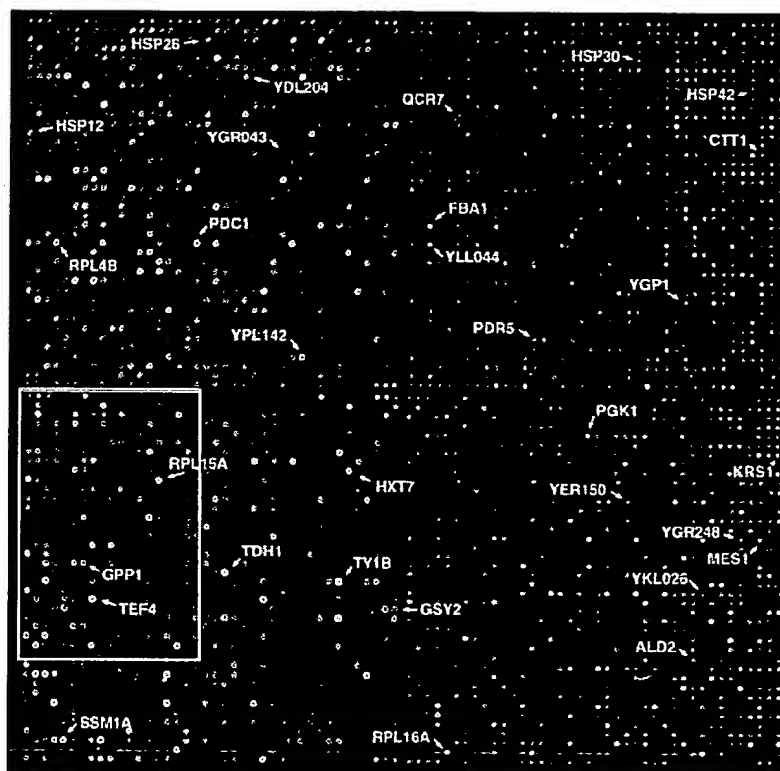


Fig. 1. Yeast genome microarray. The actual size of the microarray is 18 mm by 18 mm. The microarray was printed as described (9). This image was obtained with the same fluorescent scanning confocal microscope used to collect all the data we report (49). A fluorescently labeled cDNA probe was prepared from mRNA isolated from cells harvested shortly after inoculation (culture density of $<5 \times 10^6$ cells/ml and media glucose level of 19 g/liter) by reverse transcription in the presence of Cy3-dUTP. Similarly, a second probe was prepared from mRNA isolated from cells taken from the same culture 9.5 hours later (culture density of $\sim 2 \times 10^8$ cells/ml, with a glucose level of <0.2 g/liter) by reverse transcription in the presence of Cy5-dUTP. In this image, hybridization of the Cy3-dUTP-labeled cDNA (that is, mRNA expression at the initial timepoint) is represented as a green signal, and hybridization of Cy5-dUTP-labeled cDNA (that is, mRNA expression at 9.5 hours) is represented as a red signal. Thus, genes induced or repressed after the diauxic shift appear in this image as red and green spots, respectively. Genes expressed at roughly equal levels before and after the diauxic shift appear in this image as yellow spots.

of *HSP42*, have previously been shown to be controlled at least in part by these elements (21–24). Inspection of the sequences upstream of *HSP42* and the two uncharacterized genes shown in Fig. 5C, YK026c, a hypothetical protein with similarity to glutathione peroxidase, and YGR043c, a putative transaldolase, revealed that each of these genes also possess repeated upstream copies of the stress-responsive CCCCT motif. Of the 13 additional genes in the yeast genome that shared this expression profile [including *HSP30*, *ALD2*, *OM45*, and 10 uncharacterized ORFs (25)], nine contained one or more recognizable STRE sites in their upstream regions.

The heterotrimeric transcriptional activator complex HAP2,3,4 has been shown to be responsible for induction of several genes important for respiration (26–28). This complex binds a degenerate consensus sequence known as the CCAAT box (26). Computer analysis, using the consensus sequence TNRYTGGB (29), has suggested that a large number of genes involved in respiration may be specific targets of HAP2,3,4 (30). Indeed, a putative HAP2,3,4 binding site could be found in the sequences upstream of each of the seven cytochrome *c*-related genes that showed the greatest magnitude of induction (Fig. 5D). Of 12 additional cytochrome *c*-related genes that were induced, HAP2,3,4 binding sites were present in all but one. Significantly, we found that transcription of *HAP4* itself was induced nearly ninefold concomitant with the diauxic shift.

Control of ribosomal protein biogenesis is mainly exerted at the transcriptional level, through the presence of a common upstream-activating element (UAS_{TPG}) that is recognized by the Rap1 DNA-binding protein (31, 32). The expression profiles of seven ribosomal proteins are shown in Fig. 5F. A search of the sequences upstream of all seven genes revealed consensus Rap1-binding motifs (33). It has been suggested that declining Rap1 levels in the cell during starvation may be responsible for the decline in ribosomal protein gene expression (34). Indeed, we observed that the abundance of *RAP1* mRNA diminished by 4.4-fold, at about the time of glucose exhaustion.

Of the 149 genes that encode known or putative transcription factors, only two, *HAP4* and *SIP4*, were induced by a factor of more than threefold at the diauxic shift. *SIP4* encodes a DNA-binding transcriptional activator that has been shown to interact with Snf1, the “master regulator” of glucose repression (35). The eightfold induction of *SIP4* upon depletion of glucose strongly suggests a role in the induction of

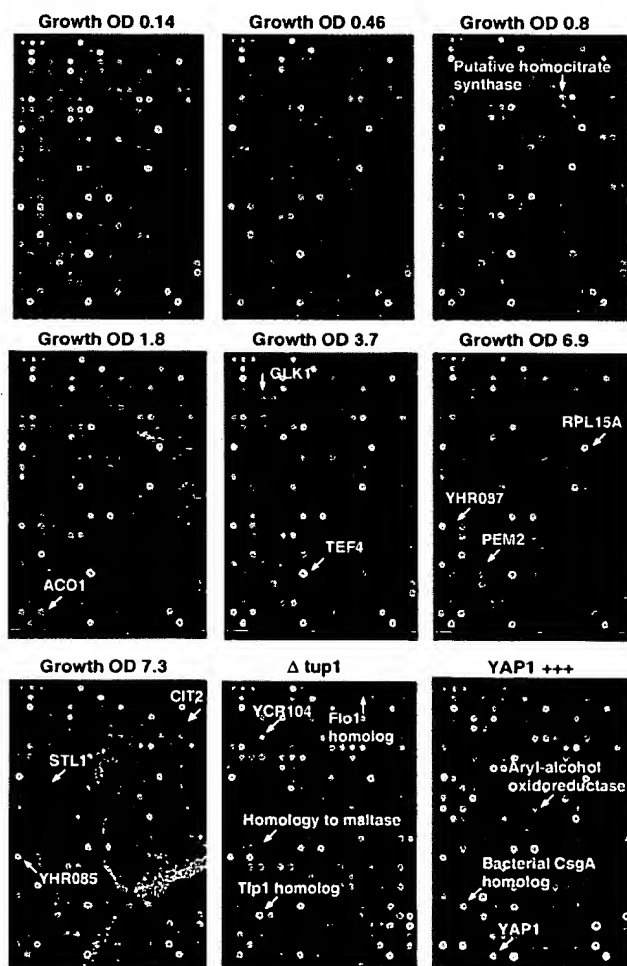
downstream genes at the diauxic shift.

Although most of the transcriptional responses that we observed were not previously known, the responses of many genes during the diauxic shift have been described. Comparison of the results we obtained by DNA microarray hybridization with previously reported results therefore provided a strong test of the sensitivity and accuracy of this approach. The expression patterns we observed for previously characterized genes showed almost perfect concordance with previously published results (36). Moreover, the differential expression measurements obtained by DNA microarray hybridization were reproducible in duplicate experiments. For example, the remarkable changes in gene expression between cells harvested immediately after inoculation and immediately after the diauxic shift (the first and sixth intervals in this time series) were measured in duplicate, independent DNA microarray hybridizations. The correlation coefficient for two complete sets of expression ratio measurements was 0.87, and for more than 95% of the genes, the expres-

sion ratios measured in these duplicate experiments differed by less than a factor of 2. However, in a few cases, there were discrepancies between our results and previous results, pointing to technical limitations that will need to be addressed as DNA microarray technology advances (37, 38). Despite the noted exceptions, the high concordance between the results we obtained in these experiments and those of previous studies provides confidence in the reliability and thoroughness of the survey.

The changes in gene expression during this diauxic shift are complex and involve integration of many kinds of information about the nutritional and metabolic state of the cell. The large number of genes whose expression is altered and the diversity of temporal expression profiles observed in this experiment highlight the challenge of understanding the underlying regulatory mechanisms. One approach to defining the contributions of individual regulatory genes to a complex program of this kind is to use DNA microarrays to identify genes whose expression is affected

Fig. 2. The section of the array indicated by the gray box in Fig. 1 is shown for each of the experiments described here. Representative genes are labeled. In each of the arrays used to analyze gene expression during the diauxic shift, red spots represent genes that were induced relative to the initial timepoint, and green spots represent genes that were repressed relative to the initial timepoint. In the arrays used to analyze the effects of the *tup1*Δ mutation and *YAP1* overexpression, red spots represent genes whose expression was increased, and green spots represent genes whose expression was decreased by the genetic modification. Note that distinct sets of genes are induced and repressed in the different experiments. The complete images of each of these arrays can be viewed on the Internet (13). Cell density as measured by optical density (OD) at 600 nm was used to measure the growth of the culture.



by mutations in each putative regulatory gene. As a test of this strategy, we analyzed the genomewide changes in gene expression that result from deletion of the *TUP1* gene. Transcriptional repression of many genes by glucose requires the DNA-binding repressor

Mig1 and is mediated by recruiting the transcriptional co-repressors Tup1 and Cyc8/Ssn6 (39). Tup1 has also been implicated in repression of oxygen-regulated, mating-type-specific, and DNA-damage-inducible genes (40).

Wild-type yeast cells and cells bearing a deletion of the *TUP1* gene (*tup1Δ*) were grown in parallel cultures in rich medium containing glucose as the carbon source. Messenger RNA was isolated from exponentially growing cells from the two populations and used to prepare cDNA labeled with Cy3 (green) and Cy5 (red), respectively (11). The labeled probes were mixed and simultaneously hybridized to the microarray. Red spots on the microarray therefore represented genes whose transcription was induced in the *tup1Δ* strain, and thus presumably repressed by Tup1 (41). A representative section of the microarray (Fig. 2, bottom middle panel) illustrates that the genes whose expression was affected by the *tup1Δ* mutation, were, in general, distinct from those induced upon glucose exhaustion [complete images of all the arrays shown in Fig. 2 are available on the Internet (13)]. Nevertheless, 34 (10%) of the genes that were induced by a factor of at least 2 after the diauxic shift were similarly induced by deletion of *TUP1*, suggesting that these genes may be subject to *TUP1*-mediated repression by glucose. For example, *SUC2*, the gene encoding invertase, and all five hexose transporter genes that were induced during the course of the diauxic shift were similarly induced, in duplicate experiments, by the deletion of *TUP1*.

The set of genes affected by Tup1 in this experiment also included α -glucosidases, the mating-type-specific genes *MFA1* and *MFA2*, and the DNA damage-inducible *RNR2* and *RNR4*, as well as genes involved in flocculation and many genes of unknown function. The hybridization signal corresponding to expression of *TUP1* itself was also severely reduced because of the (incomplete) deletion of the transcription unit in the *tup1Δ* strain, providing a positive control in the experiment (42).

Many of the transcriptional targets of Tup1 fell into sets of genes with related biochemical functions. For instance, although only about 3% of all yeast genes appeared to be *TUP1*-repressed by a factor of more than 2 in duplicate experiments under these conditions, 6 of the 13 genes that have been implicated in flocculation (15) showed a reproducible increase in expression of at least twofold when *TUP1* was deleted. Another group of related genes that appeared to be subject to *TUP1* repression encodes the serine-rich cell wall mannoproteins, such as *Tip1* and *Tir1/Srp1* which are induced by cold shock and other stresses (43), and similar, serine-poor proteins, the seripauperins (44). Messenger RNA levels for 23 of the 26 genes in this group were reproducibly elevated by at least 2.5-fold in the *tup1Δ*

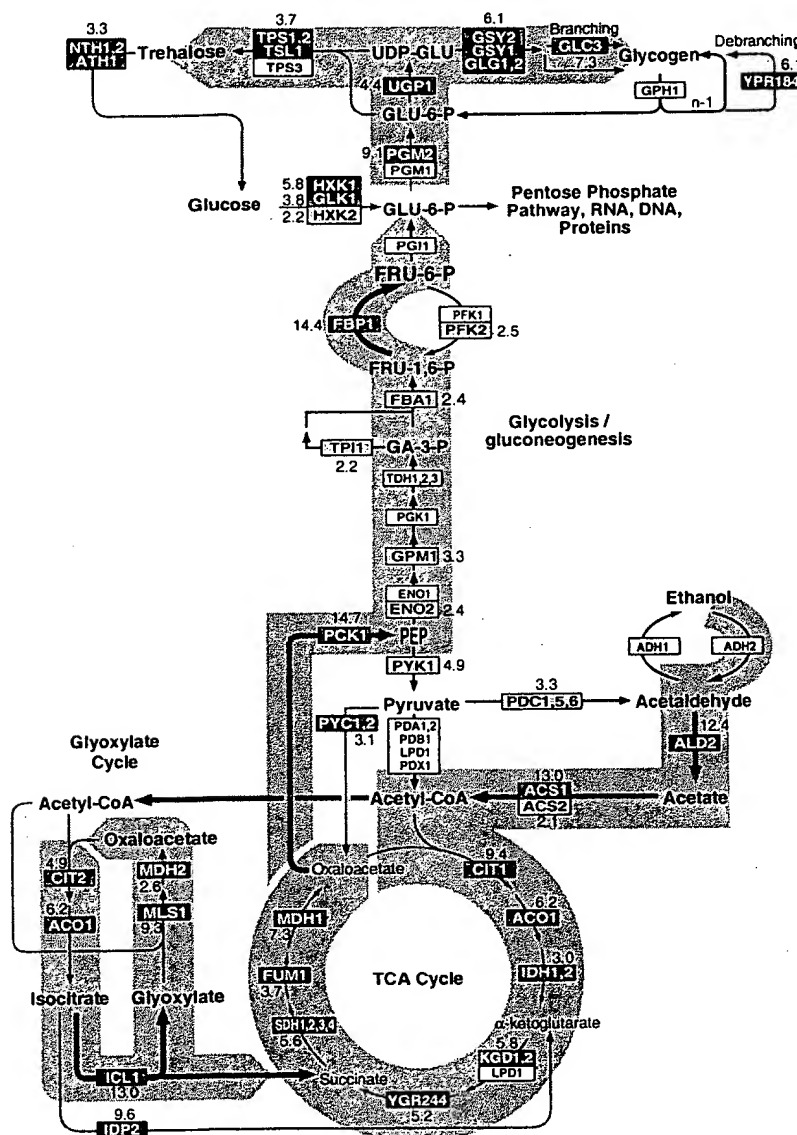


Fig. 3. Metabolic reprogramming inferred from global analysis of changes in gene expression. Only key metabolic intermediates are identified. The yeast genes encoding the enzymes that catalyze each step in this metabolic circuit are identified by name in the boxes. The genes encoding succinyl-CoA synthase and glycogen-debranching enzyme have not been explicitly identified, but the ORFs *YGR244* and *YPR184* show significant homology to known succinyl-CoA synthase and glycogen-debranching enzymes, respectively, and are therefore included in the corresponding steps in this figure. Red boxes with white lettering identify genes whose expression increases in the diauxic shift. Green boxes with dark green lettering identify genes whose expression diminishes in the diauxic shift. The magnitude of induction or repression is indicated for these genes. For multimeric enzyme complexes, such as succinate dehydrogenase, the indicated fold-induction represents an unweighted average of all the genes listed in the box. Black and white boxes indicate no significant differential expression (less than twofold). The direction of the arrows connecting reversible enzymatic steps indicate the direction of the flow of metabolic intermediates, inferred from the gene expression pattern, after the diauxic shift. Arrows representing steps catalyzed by genes whose expression was strongly induced are highlighted in red. The broad gray arrows represent major increases in the flow of metabolites after the diauxic shift, inferred from the indicated changes in gene expression.

strain, and 18 of these genes were induced by more than sevenfold when *TUP1* was deleted. In contrast, none of 83 genes that could be classified as putative regulators of the cell division cycle were induced more than twofold by deletion of *TUP1*. Thus, despite the diversity of the regulatory systems that employ Tup1, most of the genes that it regulates under these conditions fall into a limited number of distinct functional classes.

Because the microarray allows us to monitor expression of nearly every gene in yeast, we can, in principle, use this approach to identify all the transcriptional targets of a regulatory protein like Tup1. It is important to note, however, that in any single experiment of this kind we can only recognize those target genes that are normally repressed (or induced) under the conditions of the experiment. For instance, the experiment described here analyzed a MAT α strain in which *MFA1* and *MFA2*, the genes encoding the α -factor mating pheromone precursor, are normally repressed. In the isogenic *tup1 Δ* strain, these genes were inappropriately expressed, reflecting the role that Tup1 plays in their repression. Had we instead carried out this experiment with a MAT α strain (in which expression of *MFA1* and *MFA2* is not repressed), it would not have been possible to conclude anything regarding the role of Tup1 in the repression of these genes. Conversely, we cannot distinguish indirect effects of the chronic absence of Tup1 in the mutant strain from effects directly attributable to its participation in repressing the transcription of a gene.

Another simple route to modulating the activity of a regulatory factor is to overexpress the gene that encodes it. *YAP1* encodes a DNA-binding transcription factor belonging to the bZIP class of DNA-binding proteins. Overexpression of *YAP1* in yeast confers increased resistance to hydrogen peroxide, α -phenanthroline, heavy metals, and osmotic stress (45). We analyzed differential gene expression between a wild-type strain bearing a control plasmid and a strain with a plasmid expressing *YAP1* under the control of the strong *GALI-10* promoter, both grown in galactose (that is, a condition that induces *YAP1* overexpression). Complementary DNA from the control and *YAP1* overexpressing strains, labeled with Cy3 and Cy5, respectively, was prepared from mRNA isolated from the two strains and hybridized to the microarray. Thus, red spots on the array represent genes that were induced in the strain overexpressing *YAP1*.

Of the 17 genes whose mRNA levels increased by more than threefold when

YAP1 was overexpressed in this way, five bear homology to aryl-alcohol oxidoreductases (Fig. 2 and Table 1). An additional four of the genes in this set also belong to the general class of dehydrogenases/oxidoreductases. Very little is known about the role of aryl-alcohol oxidoreductases in *S. cerevisiae*, but these enzymes have been isolated from ligninolytic fungi, in which they participate in coupled redox reactions, oxidizing aromatic, and aliphatic unsaturated alcohols to aldehydes with the production of hydrogen peroxide (46, 47). The fact that a remarkable fraction of the targets identified in this experiment belong to the same small, functional group of oxidoreductases suggests that these genes

might play an important protective role during oxidative stress. Transcription of a small number of genes was reduced in the strain overexpressing *Yap1*. Interestingly, many of these genes encode sugar permeases or enzymes involved in inositol metabolism.

We searched for *Yap1*-binding sites (TTACTAA or TGACTAA) in the sequences upstream of the target genes we identified (48). About two-thirds of the genes that were induced by more than threefold upon *Yap1* overexpression had one or more binding sites within 600 bases upstream of the start codon (Table 1), suggesting that they are directly regulated by *Yap1*. The absence of canonical *Yap1*-bind-

Fig. 4. Coordinated regulation of functionally related genes. The curves represent the average induction or repression ratios for all the genes in each indicated group. The total number of genes in each group was as follows: ribosomal proteins, 112; translation elongation and initiation factors, 25; tRNA synthetases (excluding mitochondrial synthetases), 17; glycogen and trehalose synthesis and degradation, 15; cytochrome c oxidase and reductase proteins, 19; and TCA- and glyoxylate-cycle enzymes, 24.

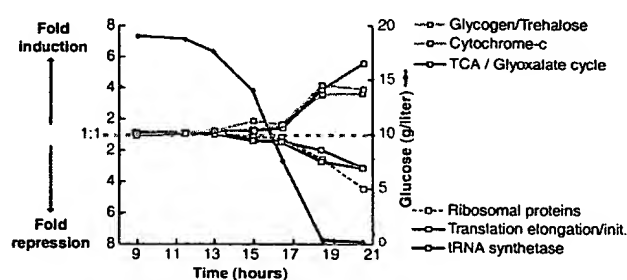


Table 1. Genes induced by *YAP1* overexpression. This list includes all the genes for which mRNA levels increased by more than twofold upon *YAP1* overexpression in both of two duplicate experiments, and for which the average increase in mRNA level in the two experiments was greater than threefold (50). Positions of the canonical *Yap1* binding sites upstream of the start codon, when present, and the average fold-increase in mRNA levels measured in the two experiments are indicated.

ORF	Distance of <i>Yap1</i> site from ATG	Gene	Description	Fold-increase
YNL331C		<i>YAP1</i>	Putative aryl-alcohol reductase	12.9
YKL071W	162–222 (5 sites)		Similarity to bacterial <i>csgA</i> protein	10.4
YML007W			Transcriptional activator involved in oxidative stress response	9.8
YFL056C	223, 242		Homology to aryl-alcohol dehydrogenases	9.0
YLL060C	98		Putative glutathione transferase	7.4
YOL165C	266		Putative aryl-alcohol dehydrogenase (NADP+)	7.0
YCR107W		<i>ATR1</i>	Putative aryl-alcohol reductase	6.5
YML116W	409		Aminotriazole and 4-nitroquinoline resistance protein	6.5
YBR008C	142, 167, 364		Homology to benomyl/methotrexate resistance protein	6.1
YCLX08C			Hypothetical protein	6.1
YJR155W			Putative aryl-alcohol dehydrogenase	6.0
YPL171C	148, 212	<i>OYE3</i>	NADPH dehydrogenase (old yellow enzyme), isoform 3	5.8
YLR460C	167, 317		Homology to hypothetical proteins YCR102c and YNL134c	4.7
YKR076W	178		Homology to hypothetical protein YMR251w	4.5
YHR179W	327	<i>OYE2</i>	NAD(P)H oxidoreductase (old yellow enzyme), isoform 1	4.1
YML131W	507		Similarity to <i>A. thaliana</i> zeta-crystallin homolog	3.7
YOL126C		<i>MDH2</i>	Malate dehydrogenase	3.3

ing sites upstream of the others may reflect an ability of Yap1 to bind sites that differ from the canonical binding sites, perhaps in cooperation with other factors, or less likely, may represent an indirect effect of Yap1 overexpression, mediated by one or more intermediary factors. Yap1 sites were found only four times in the corresponding region of an arbitrary set of 30 genes that were not differentially regulated by Yap1.

Use of a DNA microarray to characterize the transcriptional consequences of mutations affecting the activity of regulatory molecules provides a simple and powerful approach to dissection and characterization of regulatory pathways and net-

works. This strategy also has an important practical application in drug screening. Mutations in specific genes encoding candidate drug targets can serve as surrogates for the ideal chemical inhibitor or modulator of their activity. DNA microarrays can be used to define the resulting signature pattern of alterations in gene expression, and then subsequently used in an assay to screen for compounds that reproduce the desired signature pattern.

DNA microarrays provide a simple and economical way to explore gene expression patterns on a genomic scale. The hurdles to extending this approach to any other organism are minor. The equipment

required for fabricating and using DNA microarrays (9) consists of components that were chosen for their modest cost and simplicity. It was feasible for a small group to accomplish the amplification of more than 6000 genes in about 4 months and, once the amplified gene sequences were in hand, only 2 days were required to print a set of 110 microarrays of 6400 elements each. Probe preparation, hybridization, and fluorescent imaging are also simple procedures. Even conceptually simple experiments, as we described here, can yield vast amounts of information. The value of the information from each experiment of this kind will progressively increase as more is learned about the functions of each gene and as additional experiments define the global changes in gene expression in diverse other natural processes and genetic perturbations. Perhaps the greatest challenge now is to develop efficient methods for organizing, distributing, interpreting, and extracting insights from the large volumes of data these experiments will provide.

REFERENCES AND NOTES

1. M. Schena, D. Shalon, R. W. Davis, P. O. Brown, *Science* **270**, 467 (1995).
2. D. Shalon, S. J. Smith, P. O. Brown, *Genome Res.* **6**, 639 (1996).
3. D. Lashkari, *Proc. Natl. Acad. Sci. U.S.A.*, in press.
4. J. DeRisi et al., *Nature Genet.* **14**, 457 (1996).
5. D. J. Lockhart et al., *Nature Biotechnol.* **14**, 1675 (1996).
6. M. Chee et al., *Science* **274**, 610 (1996).
7. M. Johnston and M. Carlson, in *The Molecular Biology of the Yeast Saccharomyces: Gene Expression*, E. W. Jones, J. R. Pringle, J. R. Broach, Eds. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1992), p. 193.
8. Primers for each known or predicted protein coding sequence were supplied by Research Genetics. PCR was performed with the protocol supplied by Research Genetics, using genomic DNA from yeast strain S288C as a template. Each PCR product was verified by agarose gel electrophoresis and was deemed correct if the lane contained a single band of appropriate mobility. Failures were marked as such in the database. The overall success rate for a single-pass amplification of 6116 ORFs was ~94.5%.
9. Glass slides (Gold Seal) were cleaned for 2 hours in a solution of 2 N NaOH and 70% ethanol. After rinsing in distilled water, the slides were then treated with a 1:5 dilution of poly-L-lysine adhesive solution (Sigma) for 1 hour, and then dried for 5 min at 40°C in a vacuum oven. DNA samples from 100- μ l PCR reactions were purified by ethanol purification in 96-well microtiter plates. The resulting precipitates were resuspended in 3 \times standard saline citrate (SSC) and transferred to new plates for arraying. A custom-built arraying robot was used to print on a batch of 110 slides. Details of the design of the microarray are available at cmgm.stanford.edu/pbrown. After printing, the microarrays were rehydrated for 30 s in a humid chamber and then snap-dried for 2 s on a hot plate (100°C). The DNA was then ultraviolet (UV)-crosslinked to the surface by subjecting the slides to 60 mJ of energy (Stratagene Stratilinker). The rest of the poly-L-lysine surface was blocked by a 15-min incubation in a solution of 70 mM succinic anhydride dissolved in a solution consisting of 315 ml of 1-methyl-2-pyrrolidinone (Aldrich) and 35 ml of 1 M boric acid (pH 8.0). Directly after the blocking reac-

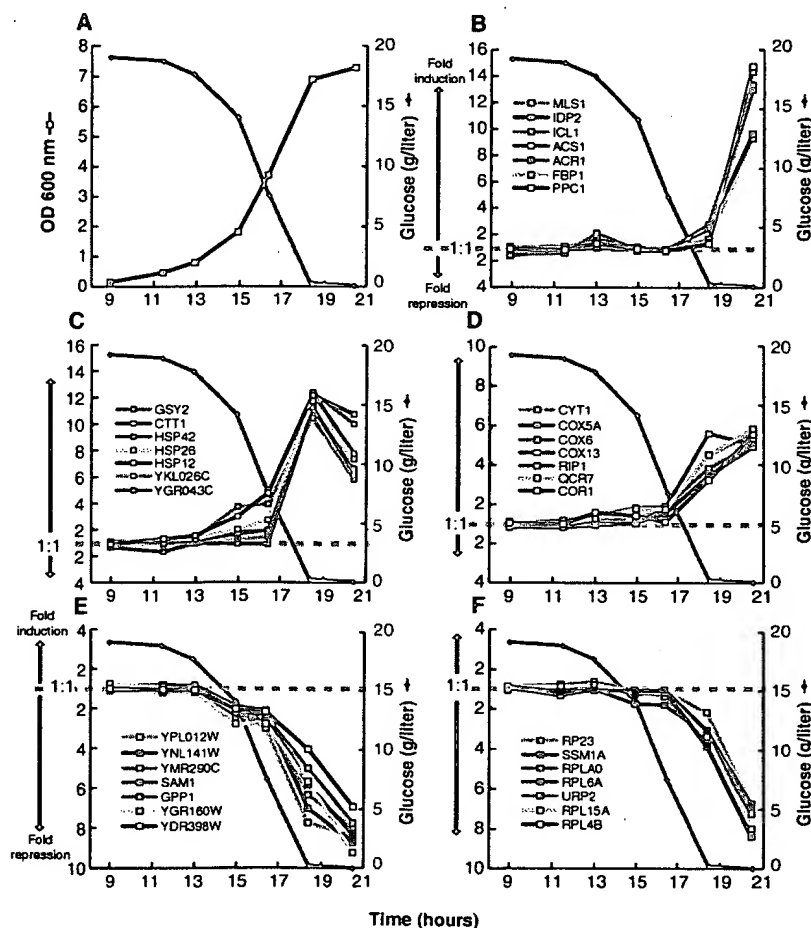


Fig. 5. Distinct temporal patterns of induction or repression help to group genes that share regulatory properties. (A) Temporal profile of the cell density, as measured by OD at 600 nm and glucose concentration in the media. (B) Seven genes exhibited a strong induction (greater than ninefold) only at the last timepoint (20.5 hours). With the exception of *IDP2*, each of these genes has a CSRE UAS. There were no additional genes observed to match this profile. (C) Seven members of a class of genes marked by early induction with a peak in mRNA levels at 18.5 hours. Each of these genes contains STRE motif repeats in their upstream promoter regions. (D) Cytochrome c oxidase and ubiquinol cytochrome c reductase genes. Marked by an induction coincident with the diauxic shift, each of these genes contains a consensus binding motif for the HAP2,3,4 protein complex. At least 17 genes shared a similar expression profile. (E) *SAM1*, *GPP1*, and several genes of unknown function are repressed before the diauxic shift, and continue to be repressed upon entry into stationary phase. (F) Ribosomal protein genes comprise a large class of genes that are repressed upon depletion of glucose. Each of the genes profiled here contains one or more RAP1-binding motifs upstream of its promoter. RAP1 is a transcriptional regulator of most ribosomal proteins.

- tion, the bound DNA was denatured by a 2-min incubation in distilled water at ~95°C. The slides were then transferred into a bath of 100% ethanol at room temperature, rinsed, and then spun dry in a clinical centrifuge. Slides were stored in a closed box at room temperature until used.
10. YPD medium (8 liters), in a 10-liter fermentation vessel, was inoculated with 2 ml of a fresh overnight culture of yeast strain DBY7286 (MATa, ura3, GAL2). The fermentor was maintained at 30°C with constant agitation and aeration. The glucose content of the media was measured with a UV test kit (Boehringer Mannheim, catalog number 716251). Cell density was measured by OD at 600-nm wavelength. Aliquots of culture were rapidly withdrawn from the fermentation vessel by peristaltic pump, spun down at room temperature, and then flash frozen with liquid nitrogen. Frozen cells were stored at -80°C.
 11. Cy3-dUTP or Cy5-dUTP (Amersham) was incorporated during reverse transcription of 1.25 µg of polyadenylated [poly(A)⁺] RNA, primed by a dT(16) oligomer. This mixture was heated to 70°C for 10 min, and then transferred to ice. A premixed solution, consisting of 200 U Superscript II (Gibco), buffer, deoxyribonucleoside triphosphates, and fluorescent nucleotides, was added to the RNA. Nucleotides were used at these final concentrations: 500 µM for dATP, dCTP, and dGTP and 200 µM for dTTP. Cy3-dUTP and Cy5-dUTP were used at a final concentration of 100 µM. The reaction was then incubated at 42°C for 2 hours. Unincorporated fluorescent nucleotides were removed by first diluting the reaction mixture with 470 µl of 10 mM tris-HCl (pH 8.0)/1 mM EDTA and then subsequently concentrating the mix to ~5 µl, using Centricon-30 microconcentrators (Amicon).
 12. Purified, labeled cDNA was resuspended in 11 µl of 3.5× SSC containing 10 µg poly(dA) and 0.3 µl of 10% SDS. Before hybridization, the solution was boiled for 2 min and then allowed to cool to room temperature. The solution was applied to the microarray under a cover slip, and the slide was placed in a custom hybridization chamber which was subsequently incubated for ~8 to 12 hours in a water bath at 62°C. Before scanning, slides were washed in 2× SSC, 0.2% SDS for 5 min, and then 0.05× SSC for 1 min. Slides were dried before scanning by centrifugation at 500 rpm in a Beckman CS-6R centrifuge.
 13. The complete data set is available on the Internet at cmgm.stanford.edu/pbrown/explore/index.html.
 14. For 95% of all the genes analyzed, the mRNA levels measured in cells harvested at the first and second interval after inoculation differed by a factor of less than 1.5. The correlation coefficient for the comparison between mRNA levels measured for each gene in these two different mRNA samples was 0.98. When duplicate mRNA preparations from the same cell sample were compared in the same way, the correlation coefficient between the expression levels measured for the two samples by comparative hybridization was 0.99.
 15. The numbers and identities of known and putative genes, and their homologies to other genes, were gathered from the following public databases: *Saccharomyces* Genome Database (genome-www.stanford.edu), Yeast Protein Database (quest7.proteome.com), and Munich Information Centre for Protein Sequences (speedy.mips.biochem.mpg.de/mips/yeast/index.html).
 16. A. Scholer and H. J. Schuller, *Mol. Cell. Biol.* 14, 3613 (1994).
 17. S. Kratzer and H. J. Schuller, *Gene* 161, 75 (1995).
 18. R. J. Haselbeck and H. L. McAlister, *J. Biol. Chem.* 268, 12116 (1993).
 19. M. Fernandez, E. Fernandez, R. Rodicio, *Mol. Gen. Genet.* 242, 727 (1994).
 20. A. Hartig et al., *Nucleic Acids Res.* 20, 5677 (1992).
 21. P. M. Martinez et al., *EMBO J.* 15, 2227 (1996).
 22. J. C. Varela, U. M. Prækelt, P. A. Meacock, R. J. Planta, W. H. Mager, *Mol. Cell. Biol.* 15, 6232 (1995).
 23. H. Ruis and C. Schuller, *Bioessays* 17, 959 (1995).
 24. J. L. Parrou, M. A. Teste, J. Francois, *Microbiology* 143, 1891 (1997).
 25. This expression profile was defined as having an induction of greater than 10-fold at 18.5 hours and less than 11-fold at 20.5 hours.
 26. S. L. Forsburg and L. Guarente, *Genes Dev.* 3, 1166 (1989).
 27. J. T. Olesen and L. Guarente, *ibid.* 4, 1714 (1990).
 28. M. Rosenkrantz, C. S. Kell, E. A. Pennell, L. J. Devenish, *Mol. Microbiol.* 13, 119 (1994).
 29. Single-letter abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr. The nucleotide codes are as follows: B—C, G, or T; N—G, A, T, or C; R—A or G; and Y—C or T.
 30. C. Fondrat and A. Kalogeropoulos, *Comput. Appl. Biosci.* 12, 363 (1996).
 31. D. Shore, *Trends Genet.* 10, 408 (1994).
 32. R. J. Planta and H. A. Raue, *ibid.* 4, 64 (1988).
 33. The degenerate consensus sequence VCYRNNC-MNH was used to search for potential RAP1-binding sites. The exact consensus, as defined by [30], is WACAYCCTACATYW, with up to three differences allowed.
 34. S. F. Neuman, S. Bhattacharya, J. R. Broach, *Mol. Cell. Biol.* 15, 3187 (1995).
 35. P. Lesage, X. Yang, M. Carlson, *ibid.* 16, 1921 (1996).
 36. For example, we observed large inductions of the genes coding for *PCK1*, *FBP1* [Z. Yin et al., *Mol. Microbiol.* 20, 751 (1996)], the central glyoxylate cycle gene *ICL1* [A. Scholer and H. J. Schuller, *Curr. Genet.* 23, 375 (1993)], and the "aerobic" isoform of acetyl-CoA synthase, *ACS1* [M. A. van den Berg et al., *J. Biol. Chem.* 271, 28953 (1996)], with concomitant down-regulation of the glycolytic-specific genes *PYK1* and *PFK2* [P. A. Moore et al., *Mol. Cell. Biol.* 11, 5330 (1991)]. Other genes not directly involved in carbon metabolism but known to be induced upon nutrient limitation include genes encoding cytosolic catalase *TCT1* [P. H. Bissinger et al., *ibid.* 9, 1309 (1989)] and several genes encoding small heat-shock proteins, such as *HSP12*, *HSP26*, and *HSP42* [I. Farkas et al., *J. Biol. Chem.* 266, 15602 (1991); U. M. Prækelt and P. A. Meacock, *Mol. Gen. Genet.* 223, 97 (1990); D. Wotton et al., *J. Biol. Chem.* 271, 2717 (1996)].
 37. The levels of induction we measured for genes that were expressed at very low levels in the uninduced state (notably, *FBP1* and *PCK1*) were generally lower than those previously reported. This discrepancy was likely due to the conservative background subtraction method we used, which generally resulted in overestimation of very low expression levels [46].
 38. Cross-hybridization of highly related sequences can also occasionally obscure changes in gene expression, an important concern where members of gene families are functionally specialized and differentially regulated. The major alcohol dehydrogenase genes, *ADH1* and *ADH2*, share 88% nucleotide identity. Reciprocal regulation of these genes is an important feature of the diauxic shift, but was not observed in this experiment, presumably because of cross-hybridization of the fluorescent cDNAs representing these two genes. Nevertheless, we were able to detect differential expression of closely related isoforms of other enzymes, such as *HXK1/HXK2* (77% identical) [P. Herrero et al., *Yeast* 11, 137 (1995)], *MLS1/DAL7* (73% identical) [20], and *PGM1/PGM2* (72% identical) [D. Oh, J. E. Hopper, *Mol. Cell. Biol.* 10, 1415 (1990)], in accord with previous studies. Use in the microarray of deliberately selected DNA sequences corresponding to the most divergent segments of homologous genes, in lieu of the complete gene sequences, should relieve this problem in many cases.
 39. F. E. Williams, U. Varanasi, R. J. Trumbly, *Mol. Cell. Biol.* 11, 3307 (1991).
 40. D. Tzamaras and K. Struhl, *Nature* 369, 758 (1994).
 41. Differences in mRNA levels between the *tup1Δ* and wild-type strain were measured in two independent experiments. The correlation coefficient between the complete sets of expression ratios measured in these duplicate experiments was 0.83. The concordance between the sets of genes that appeared to be induced was very high between the two experiments. When only the 355 genes that showed at least a twofold increase in mRNA in the *tup1Δ* strain in either of the duplicate experiments were compared, the correlation coefficient was 0.82.
 42. The *tup1Δ* mutation consists of an insertion of the LEU2 coding sequence, including a stop codon, between the ATG of *TUP1* and an Eco RI site 124 base pairs before the stop codon of the *TUP1* gene.
 43. L. R. Kowalski, K. Kondo, M. Inouye, *Mol. Microbiol.* 15, 341 (1995).
 44. M. Viswanathan, G. Muthukumar, Y. S. Cong, J. Lenard, *Gene* 148, 149 (1994).
 45. D. Hirata, K. Yano, T. Miyakawa, *Mol. Gen. Genet.* 242, 250 (1994).
 46. A. Gutierrez, L. Caramelo, A. Prieto, M. J. Martinez, A. T. Martinez, *Appl. Environ. Microbiol.* 60, 1783 (1994).
 47. A. Muheim et al., *Eur. J. Biochem.* 195, 369 (1991).
 48. J. A. Wemmie, M. S. Szczypka, D. J. Thiele, W. S. Moye-Rowley, *J. Biol. Chem.* 269, 32592 (1994).
 49. Microarrays were scanned using a custom-built scanning laser microscope built by S. Smith with software written by N. Ziv. Details concerning scanner design and construction are available at cmgm.stanford.edu/pbrown. Images were scanned at a resolution of 20 µm per pixel. A separate scan, using the appropriate excitation line, was done for each of the two fluorophores used. During the scanning process, the ratio between the signals in the two channels was calculated for several array elements containing total genomic DNA. To normalize the two channels with respect to overall intensity, we then adjusted photomultiplier and laser power settings such that the signal ratio at these elements was as close to 1.0 as possible. The combined images were analyzed with custom-written software. A bounding box, fitted to the size of the DNA spots in each quadrant, was placed over each array element. The average fluorescent intensity was calculated by summing the intensities of each pixel present in a bounding box, and then dividing by the total number of pixels. Local area background was calculated for each array element by determining the average fluorescent intensity for the lower 20% of pixel intensities. Although this method tends to underestimate the background, causing an underestimation of extreme ratios, it produces a very consistent and noise-tolerant approximation. Although the analog-to-digital board used for data collection possesses a wide dynamic range (12 bits), several signals were saturated (greater than the maximum signal intensity allowed) at the chosen settings. Therefore, extreme ratios at bright elements are generally underestimated. A signal was deemed significant if the average intensity after background subtraction was at least 2.5-fold higher than the standard deviation in the background measurements for all elements on the array.
 50. In addition to the 17 genes shown in Table 1, three additional genes were induced by an average of more than threefold in the duplicate experiments, but in one of the two experiments, the induction was less than twofold (range 1.6- to 1.9-fold).
 51. We thank H. Bennett, P. Spellman, J. Ravetto, M. Eisen, R. Pillai, B. Dunn, T. Ferea, and other members of the Brown lab for their assistance and helpful advice. We also thank S. Friend, D. Botstein, S. Smith, J. Hudson, and D. Dolginow for advice, support, and encouragement; K. Struhl and S. Chatterjee for the *Tup1* deletion strain; L. Fernandes for helpful advice on Yapi; and S. Klapholz and the reviewers for many helpful comments on the manuscript. Supported by a grant from the National Human Genome Research Institute (NHGRI) (HG00450), and by the Howard Hughes Medical Institute (HHMI). J.D.R. was supported by the HHMI and the NHGRI. V.R. was supported in part by an Institutional Training Grant in Genome Science (T32 HG00044) from the NHGRI. P.O.B. is an associate investigator of the HHMI.

5 September 1997; accepted 22 September 1997

Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships

STEVEN E. BRENNER*†‡, CYRUS CHOTHIA*, AND TIM J. P. HUBBARD§

*MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, United Kingdom; and †Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambs CB10 1SA, United Kingdom

Communicated by David R. Davies, National Institute of Diabetes, Bethesda, MD, March 16, 1998 (received for review November 12, 1997)

ABSTRACT Pairwise sequence comparison methods have been assessed using proteins whose relationships are known reliably from their structures and functions, as described in the SCOP database [Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia C. (1995) *J. Mol. Biol.* 247, 536–540]. The evaluation tested the programs BLAST [Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J. Mol. Biol.* 215, 403–410], WU-BLAST2 [Altschul, S. F. & Gish, W. (1996) *Methods Enzymol.* 266, 460–480], FASTA [Pearson, W. R. & Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. USA* 85, 2444–2448], and SSEARCH [Smith, T. F. & Waterman, M. S. (1981) *J. Mol. Biol.* 147, 195–197] and their scoring schemes. The error rate of all algorithms is greatly reduced by using statistical scores to evaluate matches rather than percentage identity or raw scores. The E-value statistical scores of SSEARCH and FASTA are reliable: the number of false positives found in our tests agrees well with the scores reported. However, the P-values reported by BLAST and WU-BLAST2 exaggerate significance by orders of magnitude. SSEARCH, FASTA $k_{\text{tup}} = 1$, and WU-BLAST2 perform best, and they are capable of detecting almost all relationships between proteins whose sequence identities are >30%. For more distantly related proteins, they do much less well; only one-half of the relationships between proteins with 20–30% identity are found. Because many homologs have low sequence similarity, most distant relationships cannot be detected by any pairwise comparison method; however, those which are identified may be used with confidence.

Sequence database searching plays a role in virtually every branch of molecular biology and is crucial for interpreting the sequences issuing forth from genome projects. Given the method's central role, it is surprising that overall and relative capabilities of different procedures are largely unknown. It is difficult to verify algorithms on sample data because this requires large data sets of proteins whose evolutionary relationships are known unambiguously and independently of the methods being evaluated. However, nearly all known homologs have been identified by sequence analysis (the method to be tested). Also, it is generally very difficult to know, in the absence of structural data, whether two proteins that lack clear sequence similarity are unrelated. This has meant that although previous evaluations have helped improve sequence comparison, they have suffered from insufficient, imperfectly characterized, or artificial test data. Assessment also has been problematic because high quality database sequence searching attempts to have both sensitivity (detection of homologs) and specificity (rejection of unrelated proteins); however, these complementary goals are linked such that increasing one causes the other to be reduced.

Sequence comparison methodologies have evolved rapidly, so no previously published tests have evaluated modern versions of programs commonly used. For example, parameters in BLAST (1) have changed, and WU-BLAST2 (2)—which produces gapped alignments—has become available. The latest version of FASTA (3) previously tested was 1.6, but the current release (version 3.0) provides fundamentally different results in the form of statistical scoring.

The previous reports also have left gaps in our knowledge. For example, there has been no published assessment of thresholds for scoring schemes more sophisticated than percentage identity. Thus, the widely discussed statistical scoring measures have never actually been evaluated on large databases of real proteins. Moreover, the different scoring schemes commonly in use have not been compared.

Beyond these issues, there is a more fundamental question: in an absolute sense, how well does pairwise sequence comparison work? That is, what fraction of homologous proteins can be detected using modern database searching methods?

In this work, we attempt to answer these questions and to overcome both of the fundamental difficulties that have hindered assessment of sequence comparison methodologies. First, we use the set of distant evolutionary relationships in the SCOP: Structural Classification of Proteins database (4), which is derived from structural and functional characteristics (5). The SCOP database provides a uniquely reliable set of homologs, which are known independently of sequence comparison. Second, we use an assessment method that jointly measures both sensitivity and specificity. This method allows straightforward comparison of different sequence searching procedures. Further, it can be used to aid interpretation of real database searches and thus provide optimal and reliable results.

Previous Assessments of Sequence Comparison. Several previous studies have examined the relative performance of different sequence comparison methods. The most encompassing analyses have been by Pearson (6, 7), who compared the three most commonly used programs. Of these, the Smith–Waterman algorithm (8) implemented in SSEARCH (3) is the oldest and slowest but the most rigorous. Modern heuristics have provided BLAST (1) the speed and convenience to make it the most popular program. Intermediate between these two is FASTA (3), which may be run in two modes offering either greater speed ($k_{\text{tup}} = 2$) or greater effectiveness ($k_{\text{tup}} = 1$). Pearson also considered different parameters for each of these programs.

To test the methods, Pearson selected two representative proteins from each of 67 protein superfamilies defined by the PIR database (9). Each was used as a query to search the database, and the matched proteins were marked as being homologous or unrelated according to their membership of PIR.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

© 1998 by The National Academy of Sciences 0027-8424/98/956073-6\$2.00/0
PNAS is available online at <http://www.pnas.org>.

Abbreviation: EPQ, errors per query.

†Present address: Department of Structural Biology, Stanford University, Fairchild Building D-109, Stanford, CA 94305-5126

‡To whom reprints requests should be addressed. e-mail: brenner@hyper.stanford.edu.

superfamilies. Pearson found that modern matrices and "In-scaling" of raw scores improve results considerably. He also reported that the rigorous Smith-Waterman algorithm worked slightly better than FASTA, which was in turn more effective than BLAST.

Very large scale analyses of matrices have been performed (10), and Henikoff and Henikoff (11) also evaluated the effectiveness of BLAST and FASTA. Their test with BLAST considered the ability to detect homologs above a predetermined score but had no penalty for methods which also reported large numbers of spurious matches. The Henikoffs searched the SWISS-PROT database (12) and used PROSITE (13) to define homologous families. Their results showed that the BLOSUM62 matrix (14) performed markedly better than the extrapolated PAM-series matrices (15), which previously had been popular.

A crucial aspect of any assessment is the data that are used to test the ability of the program to find homologs. But in Pearson's and the Henikoffs' evaluations of sequence comparison, the correct results were effectively unknown. This is because the superfamilies in PIR and PROSITE are principally created by using the same sequence comparison methods which are being evaluated. Interdependency of data and methods creates a "chicken and egg" problem, and means for example, that new methods would be penalized for correctly identifying homologs missed by older programs. For instance, immunoglobulin variable and constant domains are clearly homologous, but PIR places them in different superfamilies. The problem is widespread: each superfamily in PIR 48.00 with a structural homolog is itself homologous to an average of 1.6 other PIR superfamilies (16).

To surmount these sorts of difficulties, Sander and Schneider (17) used protein structures to evaluate sequence comparison. Rather than comparing different sequence comparison algorithms, their work focused on determining a length-dependent threshold of percentage identity, above which all proteins would be of similar structure. A result of this analysis was the HSSP equation; it states that proteins with 25% identity over 80 residues will have similar structures, whereas shorter alignments require higher identity. (Other studies also have used structures (18–20), but these focused on a small number of model proteins and were principally oriented toward evaluating alignment accuracy rather than homology detection.)

A general solution to the problem of scoring comes from statistical measures (i.e., E-values and P-values) based on the extreme value distribution (21). Extreme value scoring was implemented analytically in the BLAST program using the Karlin and Altschul statistics (22, 23) and empirical approaches have been recently added to FASTA and SSEARCH. In addition to being heralded as a reliable means of recognizing significantly similar proteins (24, 25), the mathematical tractability of statistical scores "is a crucial feature of the BLAST algorithm" (1). The validity of this scoring procedure has been tested analytically and empirically (see ref. 2 and references in ref. 24). However, all large empirical tests used random sequences that may lack the subtle structure found within biological sequences (26, 27) and obviously do not contain any real homologs. Thus, although many researchers have suggested that statistical scores be used to rank matches (24, 25, 28), there have been no large rigorous experiments on biological data to determine the degree to which such rankings are superior.

A Database for Testing Homology Detection. Since the discovery that the structures of hemoglobin and myoglobin are very similar though their sequences are not (29), it has been apparent that comparing structures is a more powerful (if less convenient) way to recognize distant evolutionary relationships than comparing sequences. If two proteins show a high degree of similarity in their structural details and function, it

is very probable that they have an evolutionary relationship though their sequence similarity may be low.

The recent growth of protein structure information combined with the comprehensive evolutionary classification in the SCOP database (4, 5) have allowed us to overcome previous limitations. With these data, we can evaluate the performance of sequence comparison methods on real protein sequences whose relationships are known confidently. The SCOP database uses structural information to recognize distant homologs, the large majority of which can be determined unambiguously. These superfamilies, such as the globins or the immunoglobulins, would be recognized as related by the vast majority of the biological community despite the lack of high sequence similarity.

From SCOP, we extracted the sequences of domains of proteins in the Protein Data Bank (PDB) (30) and created two databases. One (PDB90D-B) has domains, which were all <90% identical to any other, whereas (PDB40D-B) had those <40% identical. The databases were created by first sorting all protein domains in SCOP by their quality and making a list. The highest quality domain was selected for inclusion in the database and removed from the list. Also removed from the list (and discarded) were all other domains above the threshold level of identity to the selected domain. This process was repeated until the list was empty. The PDB40D-B database contains 1,323 domains, which have 9,044 ordered pairs of distant relationships, or $\approx 0.5\%$ of the total 1,749,006 ordered pairs. In PDB90D-B, the 2,079 domains have 53,988 relationships, representing 1.2% of all pairs. Low complexity regions of sequence can achieve spurious high scores, so these were masked in both databases by processing with the SEG program (27) using recommended parameters: 12 1.8 2.0. The databases used in this paper are available from <http://sss.stanford.edu/sss/>, and databases derived from the current version of SCOP may be found at <http://scop.mrc-lmb.cam.ac.uk/scop/>.

Analyses from both databases were generally consistent, but PDB40D-B focuses on distantly related proteins and reduces the heavy overrepresentation in the PDB of a small number of families (31, 32), whereas PDB90D-B (with more sequences) improves evaluations of statistics. Except where noted otherwise, the distant homolog results here are from PDB40D-B. Although the precise numbers reported here are specific to the structural domain databases used, we expect the trends to be general.

Assessment Data and Procedure. Our assessment of sequence comparison may be divided into four different major categories of tests. First, using just a single sequence comparison algorithm at a time, we evaluated the effectiveness of different scoring schemes. Second, we assessed the reliability of scoring procedures, including an evaluation of the validity of statistical scoring. Third, we compared sequence comparison algorithms (using the optimal scoring scheme) to determine their relative performance. Fourth, we examined the distribution of homologs and considered the power of pairwise sequence comparison to recognize them. All of the analyses used the databases of structurally identified homologs and a new assessment criterion.

The analyses tested BLAST (1), version 1.4.9MP, and WU-BLAST2 (2), version 2.0a13MP. Also assessed was the FASTA package, version 3.0t76 (3), which provided FASTA and the SSEARCH implementation of Smith-Waterman (8). For SSEARCH and FASTA, we used BLOSUM45 with gap penalties $-12/-1$ (7, 16). The default parameters and matrix (BLOSUM62) were used for BLAST and WU-BLAST2.

The "Coverage Vs. Error" Plot. To test a particular protocol (comprising a program and scoring scheme), each sequence from the database was used as a query to search the database. This yielded ordered pairs of query and target sequences with associated scores, which were sorted, on the basis of their scores, from best to worst. The ideal method would have

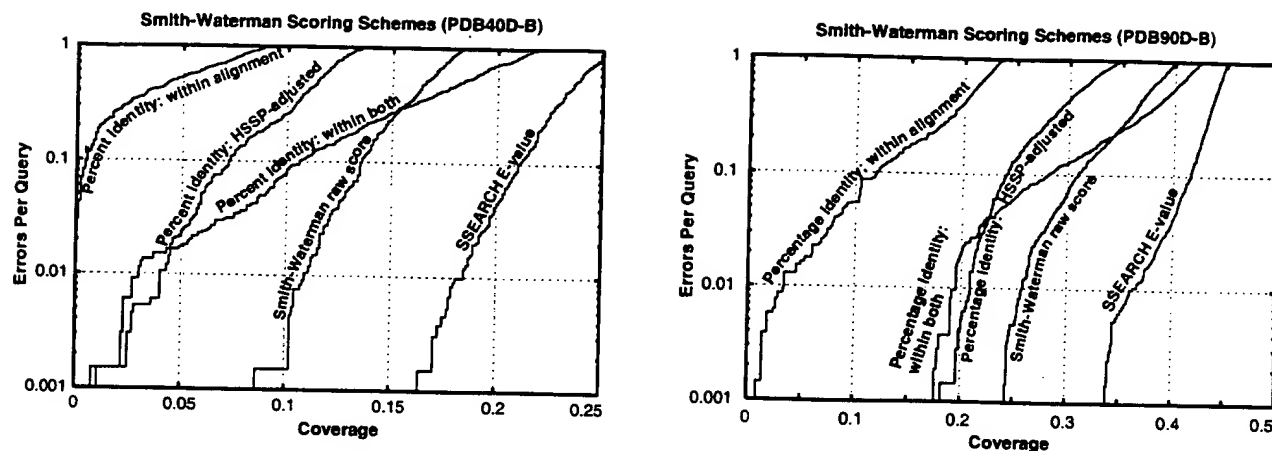


FIG. 1. Coverage vs. error plots of different scoring schemes for SSEARCH Smith-Waterman. (A) Analysis of PDB40D-B database. (B) Analysis of PDB90D-B database. All of the proteins in the database were compared with each other using the SSEARCH program. The results of this single set of comparisons were considered using five different scoring schemes and assessed. The graphs show the coverage and errors per query (EPQ) for statistical scores, raw scores, and three measures using percentage identity. In the coverage vs. error plot, the x axis indicates the fraction of all homologs in the database (known from structure) which have been detected. Precisely, it is the number of detected pairs of proteins with the same fold divided by the total number of pairs from a common superfamily. PDB40D-B contains a total of 9,044 homologs, so a score of 10% indicates identification of 904 relationships. The y axis reports the number of EPQ. Because there are 1,323 queries made in the PDB40D-B all-vs.-all comparison, 13 errors corresponds to 0.01, or 1% EPQ. The y axis is presented on a log scale to show results over the widely varying degrees of accuracy which may be desired. The scores that correspond to the levels of EPQ and coverage are shown in Fig. 4 and Table 1. The graph demonstrates the trade-off between sensitivity and selectivity. As more homologs are found (moving to the right), more errors are made (moving up). The ideal method would be in the lower right corner of the graph, which corresponds to identifying many evolutionary relationships without selecting unrelated proteins. Three measures of percentage identity are plotted. Percentage identity within alignment is the degree of identity within the aligned region of the proteins, without consideration of the alignment length. Percentage identity within both is the number of identical residues in the aligned region as a percentage of the average length of the query and target proteins. The HSP equation (17) is $H = 290.15l^{-0.562}$ where l is length for $10 < l < 80$; $H = 24.7$ for $l < 10$; $H = 24.7$ for $l > 80$. The percentage identity HSP-adjusted score is the percent identity within the alignment minus H . Smith-Waterman raw scores and E-values were taken directly from the sequence comparison program.

perfect separation, with all of the homologs at the top of the list and unrelated proteins below. In practice, perfect separation is impossible to achieve so instead one is interested in drawing a threshold above which there are the largest number of related pairs of sequences consistent with an acceptable error rate.

Our procedure involved measuring the coverage and error for every threshold. Coverage was defined as the fraction of structurally determined homologs that have scores above the selected threshold; this reflects the sensitivity of a method. Errors per query (EPQ), an indicator of selectivity, is the number of nonhomologous pairs above the threshold divided by the number of queries. Graphs of these data, called coverage vs. error plots, were devised to understand how

protocols compare at different levels of accuracy. These graphs share effectively all of the beneficial features of Receiver Operating Characteristic (ROC) plots (33, 34) but better represent the high degrees of accuracy required in sequence comparison and the huge background of nonhomologs.

This assessment procedure is directly relevant to practical sequence database searching, for it provides precisely the information necessary to perform a reliable sequence database search. The EPQ measure places a premium on score consistency; that is, it requires scores to be comparable for different queries. Consistency is an aspect which has been largely

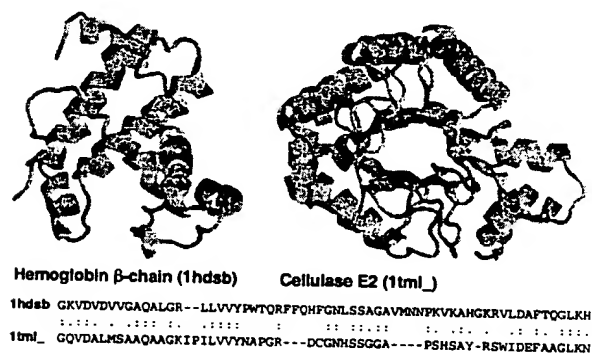


FIG. 2. Unrelated proteins with high percentage identity. Hemoglobin β -chain (PDB code 1hds chain b, ref. 38, Left) and cellulase E2 (PDB code 1tml, ref. 39, Right) have 39% identity over 64 residues, a level which is often believed to be indicative of homology. Despite this high degree of identity, their structures strongly suggest that these proteins are not related. Appropriately, neither the raw alignment score of 85 nor the E-value of 1.3 is significant. Proteins rendered by RASMOL (40).

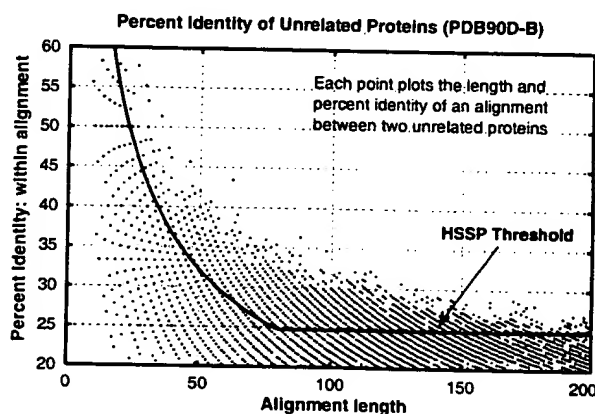


FIG. 3. Length and percentage identity of alignments of unrelated proteins in PDB90D-B: Each pair of nonhomologous proteins found with SSEARCH is plotted as a point whose position indicates the length and the percentage identity within the alignment. Because alignment length and percentage identity are quantized, many pairs of proteins may have exactly the same alignment length and percentage identity. The line shows the HSP threshold (though it is intended to be applied with a different matrix and parameters).

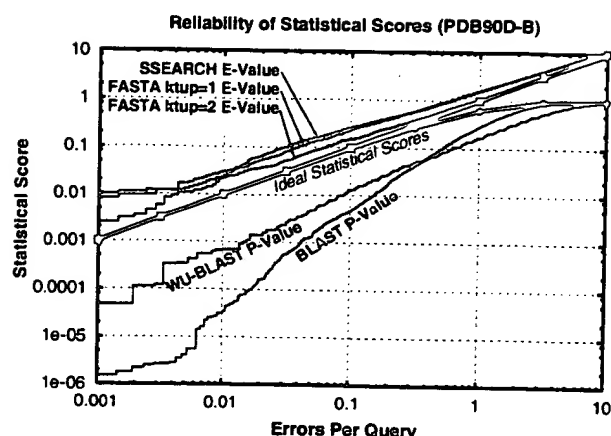


FIG. 4. Reliability of statistical scores in PDB90D-B: Each line shows the relationship between reported statistical score and actual error rate for a different program. E-values are reported for SSEARCH and FASTA, whereas P-values are shown for BLAST and WU-BLAST2. If the scoring were perfect, then the number of errors per query and the E-values would be the same, as indicated by the upper bold line. (P-values should be the same as EPQ for small numbers, and diverges at higher values, as indicated by the lower bold line.) E-values from SSEARCH and FASTA are shown to have good agreement with EPQ but underestimate the significance slightly. BLAST and WU-BLAST2 are overconfident, with the degree of exaggeration dependent upon the score. The results for PDB40D-B were similar to those for PDB90D-B despite the difference in number of homologs detected. This graph could be used to roughly calibrate the reliability of a given statistical score.

ignored in previous tests but is essential for the straightforward or automatic interpretation of sequence comparison results. Further, it provides a clear indication of the confidence that should be ascribed to each match. Indeed, the EPQ measure should approximate the expectation value reported by database searching programs, if the programs' estimates are accurate.

The Performance of Scoring Schemes. All of the programs tested could provide three fundamental types of scores. The first score is the percentage identity, which may be computed in several ways based on either the length of the alignment or the lengths of the sequences. The second is a "raw" or "Smith-Waterman" score, which is the measure optimized by the Smith-Waterman algorithm and is computed by summing the substitution matrix scores for each position in the alignment and subtracting gap penalties. In BLAST, a measure

related to this score is scaled into bits. Third is a statistical score based on the extreme value distribution. These results are summarized in Fig. 1.

Sequence Identity. Though it has been long established that percentage identity is a poor measure (35), there is a common rule-of-thumb stating that 30% identity signifies homology. Moreover, publications have indicated that 25% identity can be used as a threshold (17, 36). We find that these thresholds, originally derived years ago, are not supported by present results. As databases have grown, so have the possibilities for chance alignments with high identity; thus, the reported cutoffs lead to frequent errors. Fig. 2 shows one of the many pairs of proteins with very different structures that nonetheless have high levels of identity over considerable aligned regions. Despite the high identity, the raw and the statistical scores for such incorrect matches are typically not significant. The principal reasons percentage identity does so poorly seem to be that it ignores information about gaps and about the conservative or radical nature of residue substitutions.

From the PDB90D-B analysis in Fig. 3, we learn that 30% identity is a reliable threshold for this database only for sequence alignments of at least 150 residues. Because one unrelated pair of proteins has 43.5% identity over 62 residues, it is probably necessary for alignments to be at least 70 residues in length before 40% is a reasonable threshold, for a database of this particular size and composition.

At a given reliability, scores based on percentage identity detect just a fraction of the distant homologs found by statistical scoring. If one measures the percentage identity in the aligned regions without consideration of alignment length, then a negligible number of distant homologs are detected. Use of the HSP equation improves the value of percentage identity, but even this measure can find only 4% of all known homologs at 1% EPQ. In short, percentage identity discards most of the information measured in a sequence comparison.

Raw Scores. Smith-Waterman raw scores perform better than percentage identity (Fig. 1), but ln-scaling (7) provided no notable benefit in our analysis. It is necessary to be very precise when using either raw or bit scores because a 20% change in cutoff score could yield a tenfold difference in EPQ. However, it is difficult to choose appropriate thresholds because the reliability of a bit score depends on the lengths of the proteins matched and the size of the database. Raw score thresholds also are affected by matrix and gap parameters.

Statistical Scores. Statistical scores were introduced partly to overcome the problems that arise from raw scores. This scoring scheme provides the best discrimination between homologous proteins and those which are unrelated. Most

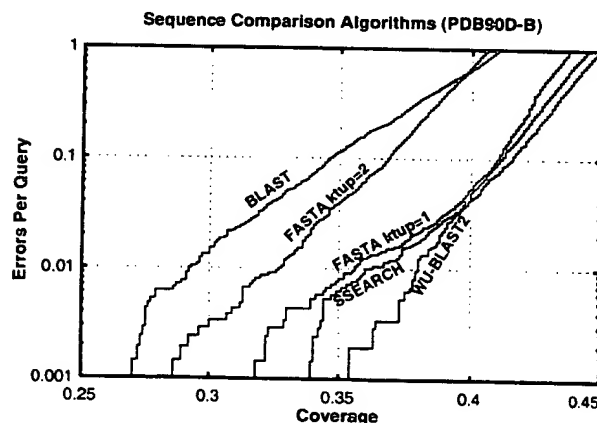
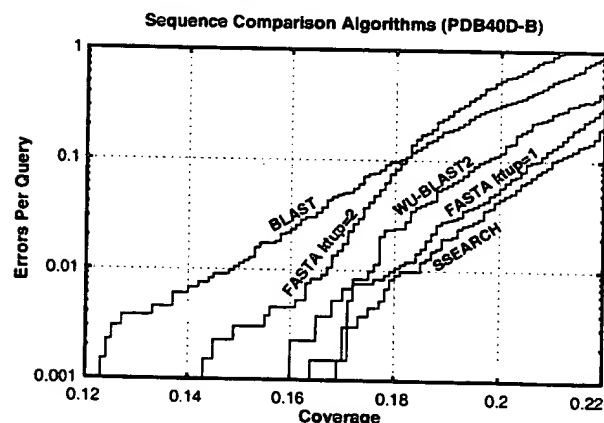


FIG. 5. Coverage vs. error plots of different sequence comparison methods: Five different sequence comparison methods are evaluated, each using statistical scores (E- or P-values). (A) PDB40D-B database. In this analysis, the best method is the slow SSEARCH, which finds 18% of relationships at 1% EPQ. FASTA ktup = 1 and WU-BLAST2 are almost as good. (B) PDB90D-B database. The quick WU-BLAST2 program provides the best coverage at 1% EPQ on this database, although at higher levels of error it becomes slightly worse than FASTA ktup = 1 and SSEARCH.

likely, its power can be attributed to its incorporation of more information than any other measure; it takes account of the full substitution and gap data (like raw scores) but also has details about the sequence lengths and composition and is scaled appropriately.

We find that statistical scores are not only powerful, but also easy to interpret. SSEARCH and FASTA show close agreement between statistical scores and actual number of errors per query (Fig. 4). The expectation value score gives a good, slightly conservative estimate of the chances of the two sequences being found at random in a given query. Thus, an E-value of 0.01 indicates that roughly one pair of nonhomologs of this similarity should be found in every 100 different queries. Neither raw scores nor percentage identity can be interpreted in this way, and these results validate the suitability of the extreme value distribution for describing the scores from a database search.

The P-values from BLAST also should be directly interpretable but were found to overstate significance by more than two orders of magnitude for 1% EPQ for this database. Nonetheless, these results strongly suggest that the analytic theory is fundamentally appropriate. WU-BLAST2 scores were more reliable than those from BLAST, but also exaggerate expected confidence by more than an order of magnitude at 1% EPQ.

Overall Detection of Homologs and Comparison of Algorithms. The results in Fig. 5A and Table 1 show that pairwise sequence comparison is capable of identifying only a small fraction of the homologous pairs of sequences in PDB40D-B. Even SSEARCH with E-values, the best protocol tested, could find only 18% of all relationships at a 1% EPQ. BLAST, which identifies 15%, was the worst performer, whereas FASTA $k_{\text{up}} = 1$ is nearly as effective as SSEARCH. FASTA $k_{\text{up}} = 2$ and WU-BLAST2 are intermediate in their ability to detect homologs. Comparison of different algorithms indicates that those capable of identifying more homologs are generally slower. SSEARCH is 25 times slower than BLAST and 6.5 times slower than FASTA $k_{\text{up}} = 1$. WU-BLAST2 is slightly faster than FASTA $k_{\text{up}} = 2$, but the latter has more interpretable scores.

In PDB90D-B, where there are many close relationships, the best method can identify only 38% of structurally known homologs (Fig. 5B). The method which finds that many relationships is WU-BLAST2. Consequently, we infer that the differences between FASTA $k_{\text{up}} = 1$, SSEARCH, and WU-BLAST2 programs are unlikely to be significant when compared with variation in database composition and scoring reliability.

Fig. 6 helps to explain why most distant homologs cannot be found by sequence comparison: a great many such relationships have no more sequence identity than would be expected by chance. SSEARCH with E-values can recognize >90% of the homologous pairs with 30–40% identity. In this region, there are 30 pairs of homologous proteins that do not have significant E-values, but 26 of these involve sequences with <50 residues. Of sequences having 25–30% identity, 75% are identified by SSEARCH E-values. However, although the number of homologs grows at lower levels of identity, the detection falls off sharply: only 40% of homologs with 20–25% identity

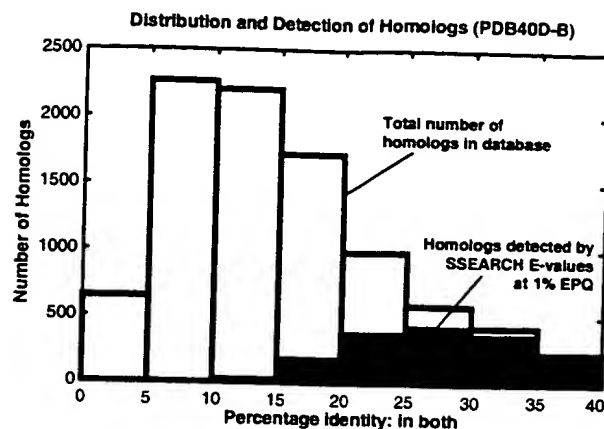


FIG. 6. Distribution and detection of homologs in PDB40D-B. Bars show the distribution of homologous pairs PDB40D-B according to their identity (using the measure of identity in both). Filled regions indicate the number of these pairs found by the best database searching method (SSEARCH with E-values) at 1% EPQ. The PDB40D-B database contains proteins with <40% identity, and as shown on this graph, most structurally identified homologs in the database have diverged extremely far in sequence and have <20% identity. Note that the alignments may be inaccurate, especially at low levels of identity. Filled regions show that SSEARCH can identify most relationships that have 25% or more identity, but its detection wanes sharply below 25%. Consequently, the great sequence divergence of most structurally identified evolutionary relationships effectively defeats the ability of pairwise sequence comparison to detect them.

are detected and only 10% of those with 15–20% can be found. These results show that statistical scores can find related proteins whose identity is remarkably low; however, the power of the method is restricted by the great divergence of many protein sequences.

After completion of this work, a new version of pairwise BLAST was released: BLASTGP (37). It supports gapped alignments, like WU-BLAST2, and dispenses with sum statistics. Our initial tests on BLASTGP using default parameters show that its E-values are reliable and that its overall detection of homologs was substantially better than that of ungapped BLAST, but not quite equal to that of WU-BLAST2.

CONCLUSION

The general consensus amongst experts (see refs. 7, 24, 25, 27 and references therein) suggests that the most effective sequence searches are made by (i) using a large current database in which the protein sequences have been complexity masked and (ii) using statistical scores to interpret the results. Our experiments fully support this view.

Our results also suggest two further points. First, the E-values reported by FASTA and SSEARCH give fairly accurate estimates of the significance of each match, but the P-values provided by BLAST and WU-BLAST2 underestimate the true

Table 1. Summary of sequence comparison methods with PDB40D-B

Method	Relative Time*	1% EPQ Cutoff	Coverage at 1% EPQ
SSEARCH % identity: within alignment	25.5	>70%	<0.1
SSEARCH % identity: within both	25.5	34%	3.0
SSEARCH % identity: HSSP-scaled	25.5	35% (HSSP + 9.8)	4.0
SSEARCH Smith–Waterman raw scores	25.5	142	10.5
SSEARCH E-values	25.5	0.03	18.4
FASTA $k_{\text{up}} = 1$ E-values	3.9	0.03	17.9
FASTA $k_{\text{up}} = 2$ E-values	1.4	0.03	16.7
WU-BLAST2 P-values	1.1	0.003	17.5
BLAST P-values	1.0	0.00016	14.8

*Times are from large database searches with genome proteins.

extent of errors. Second, SSEARCH, WU-BLAST2, and FASTA ktup = 1 perform best, though BLAST and FASTA ktup = 2 detect most of the relationships found by the best procedures and are appropriate for rapid initial searches.

The homologous proteins that are found by sequence comparison can be distinguished with high reliability from the huge number of unrelated pairs. However, even the best database searching procedures tested fail to find the large majority of distant evolutionary relationships at an acceptable error rate. Thus, if the procedures assessed here fail to find a reliable match, it does not imply that the sequence is unique; rather, it indicates that any relatives it might have are distant ones.**

**Additional and updated information about this work, including supplementary figures, may be found at <http://sss.stanford.edu/sss/>.

The authors are grateful to Drs. A. G. Murzin, M. Levitt, S. R. Eddy, and G. Mitchison for valuable discussion. S.E.B. was principally supported by a St. John's College (Cambridge, UK) Benefactors' Scholarship and by the American Friends of Cambridge University. S.E.B. dedicates his contribution to the memory of Rabbi Albert T. and Clara S. Bilgray.

- Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J. Mol. Biol.* **215**, 403–410.
- Altschul, S. F. & Gish, W. (1996) *Methods Enzymol.* **266**, 460–480.
- Pearson, W. R. & Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 2444–2448.
- Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995) *J. Mol. Biol.* **247**, 536–540.
- Brenner, S. E., Chothia, C., Hubbard, T. J. P. & Murzin, A. G. (1996) *Methods Enzymol.* **266**, 635–643.
- Pearson, W. R. (1991) *Genomics* **11**, 635–650.
- Pearson, W. R. (1995) *Protein Sci.* **4**, 1145–1160.
- Smith, T. F. & Waterman, M. S. (1981) *J. Mol. Biol.* **147**, 195–197.
- George, D. G., Hunt, L. T. & Barker, W. C. (1996) *Methods Enzymol.* **266**, 41–59.
- Vogt, G., Etzold, T. & Argos, P. (1995) *J. Mol. Biol.* **249**, 816–831.
- Henikoff, S. & Henikoff, J. G. (1993) *Proteins* **17**, 49–61.
- Bairoch, A. & Apweiler, R. (1996) *Nucleic Acids Res.* **24**, 21–25.
- Bairoch, A., Bucher, P. & Hofmann, K. (1996) *Nucleic Acids Res.* **24**, 189–196.
- Henikoff, S. & Henikoff, J. G. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 10915–10919.
- Dayhoff, M., Schwartz, R. M. & Orcutt, B. C. (1978) in *Atlas of Protein Sequence and Structure*, ed. Dayhoff, M. (National Bio-medical Research Foundation, Silver Spring, MD). Vol. 5, Suppl. 3, pp. 345–352.
- Brenner, S. E. (1996) Ph.D. thesis. (University of Cambridge, UK).
- Sander, C. & Schneider, R. (1991) *Proteins* **9**, 56–68.
- Johnson, M. S. & Overington, J. P. (1993) *J. Mol. Biol.* **233**, 716–738.
- Barton, G. J. & Sternberg, M. J. E. (1987) *Protein Eng.* **1**, 89–94.
- Lesk, A. M., Levitt, M. & Chothia, C. (1986) *Protein Eng.* **1**, 77–78.
- Arratia, R., Gordon, L. & M. W. (1986) *Ann. Stat.* **14**, 971–993.
- Karlin, S. & Altschul, S. F. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 2264–2268.
- Karlin, S. & Altschul, S. F. (1993) *Proc. Natl. Acad. Sci. USA* **90**, 5873–5877.
- Altschul, S. F., Boguski, M. S., Gish, W. & Wootton, J. C. (1994) *Nat. Genet.* **6**, 119–129.
- Pearson, W. R. (1996) *Methods Enzymol.* **266**, 227–258.
- Lipman, D. J., Wilbur, W. J., Smith, T. F. & Waterman, M. S. (1984) *Nucleic Acids Res.* **12**, 215–226.
- Wootton, J. C. & Federhen, S. (1996) *Methods Enzymol.* **266**, 554–571.
- Waterman, M. S. & Vingron, M. (1994) *Stat. Science* **9**, 367–381.
- Perutz, M. F., Kendrew, J. C. & Watson, H. C. (1965) *J. Mol. Biol.* **13**, 669–678.
- Abola, E. E., Bernstein, F. C., Bryant, S. H., Koetzle, T. F. & Weng, J. (1987) in *Crystallographic Databases: Information Content, Software Systems, Scientific Applications*, eds. Allen, F. H., Bergerhoff, G. & Sievers, R. (Data Comm. Intl. Union Crystallogr., Cambridge, UK), pp. 107–132.
- Brenner, S. E., Chothia, C. & Hubbard, T. J. P. (1997) *Curr. Opin. Struct. Biol.* **7**, 369–376.
- Orengo, C., Michie, A., Jones, S., Jones, D. T., Swindells, M. B. & Thornton, J. (1997) *Structure (London)* **5**, 1093–1108.
- Zweig, M. H. & Campbell, G. (1993) *Clin. Chem.* **39**, 561–577.
- Gribskov, M. & Robinson, N. L. (1996) *Comput. Chem.* **20**, 25–33.
- Fitch, W. M. (1966) *J. Mol. Biol.* **16**, 9–16.
- Chung, S. Y. & Subbiah, S. (1996) *Structure (London)* **4**, 1123–1127.
- Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acids Res.* **25**, 3389–3402.
- Girling, R., Schmidt, W., Jr, Houston, T., Amma, E. & Huisman, T. (1979) *J. Mol. Biol.* **131**, 417–433.
- Spezio, M., Wilson, D. & Karplus, P. (1993) *Biochemistry* **32**, 9906–9916.
- Sayle, R. A. & Milner-White, E. J. (1995) *Trends Biochem. Sci.* **20**, 374–376.

N. Leigh Anderson
Ricardo Esquer-Blasco
Jean-Paul Hofmann
Norman G. Anderson

Large Scale Biology Corporation,
Rockville, MD

A two-dimensional gel database of rat liver proteins useful in gene regulation and drug effects studies

A standard two-dimensional (2-D) protein map of Fischer 344 rat liver (F344MST3) is presented, with a tabular listing of more than 1200 protein species. Sodium dodecyl sulfate (SDS) molecular mass and isoelectric point have been established, based on positions of numerous internal standards. This map has been used to connect and compare hundreds of 2-D gels of rat liver samples from a variety of studies, and forms the nucleus of an expanding database describing rat liver proteins and their regulation by various drugs and toxic agents. An example of such a study, involving regulation of cholesterol synthesis by cholesterol-lowering drugs and a high-cholesterol diet, is presented. Since the map has been obtained with a widely used and highly reproducible 2-D gel system (the Iso-Dalt[®] system), it can be directly related to an expanding body of work in other laboratories.

Contents

1 Introduction.....	907
2 Material and methods	908
2.1 Sample preparation.....	908
2.2 Two-dimensional electrophoresis	909
2.3 Staining.....	909
2.4 Positional standardization	909
2.5 Computer analysis.....	909
2.6 Graphical data output	910
2.7 Experiment LSBC04.....	910
3 Results and discussion.....	910
3.1 The rat liver protein 2-D map.....	910
3.2 Carbamylated charge standards computed pI's and molecular mass standardization	911
3.3 An example of rat liver gene regulation: Cholesterol metabolism	911
3.3.1 MSN 413 (putative cytosolic HMG-CoA synthase) and sets of spots regulated coordinately or inversely	911
3.3.2 MSN 235 and coregulated spots.....	912
3.3.3 An example of an anti-synergistic effect	912
3.3.4 Complexity of the cholesterol synthesis pathway	912
4 Conclusions	912
5 References	912
6 Addendum 1: Figures 1-13.....	914
Addendum 2: Tables 1-4.....	923
Table 1. Master table of proteins in rat liver database	923
Table 2. Table of some identified proteins	928
Table 3. Computed pI's of two sets of carbamylated protein standards: rabbit muscle CPK and human Hb.....	929
Table 4. Computed pI's of some known proteins related to measured CPK pI's.....	930

1 Introduction

High-resolution two-dimensional electrophoresis of proteins, introduced in 1975 by O'Farrell and others [1-4], has been used over the ensuing 16 years to examine a wide variety of biological systems, the results appearing in more than 5000 published papers. With the advent of computerized systems for analyzing two-dimensional (2-D) gel images and constructing spot databases, it is also possible to plan and assemble integrated bodies of information describing the appearance and regulation of thousands of protein gene products [5, 6]. Creating such databases involves amassing and organizing quantitative data from thousands of 2-D gels, and requires a substantial commitment in technology and resources.

Given the long-term effort required to develop a protein database, the choice of a biological system takes on considerable importance. While *in vitro* systems are ideal for answering many experimental questions, especially in cancer research and genetics, our experience with cell cultures and tissue samples suggests that some *in vivo* approaches could have major advantages. In particular, we have noticed that liver tissue samples from rats and mice appear to show greater quantitative reproducibility (in terms of individual protein expression) than replicate cell cultures. This is perhaps a natural result of the homeostasis maintained in a complete animal vs. the well-known variability of cell cultures, the latter due principally to differences in reagents (e.g., fetal bovine serum), conditions (e.g., pH) and genetic "evolution" of cell lines while in culture. It is also more difficult to generate adequate amounts of protein from cell culture systems (particularly with attached cells), forcing the investigator to resort to radioisotope-based or silver-based stain-detection methods. While these methods are more sensitive (sometimes much more sensitive) than the Coomassie Brilliant Blue (CBB) stain typically used for protein detection in "large" protein samples, they are generally more variable, more labor-intensive and, in the case of radiographic methods, may generate highly "noisy" images, due to the properties of the films used. By contrast, large protein samples can easily be prepared from liver using urea/Nonidet P-40 (NP-40) solubilization and stained with CBB, which has the advantage of being easily reproducible [8]. Finally, there remains the question of the "truthfulness" of many *in vitro* systems as compared to their *in vivo* analogs; how great are the changes caused by the introduction into a cul-

Correspondence: Dr. N. Leigh Anderson, Large Scale Biology Corporation, 9620 Medical Center Drive, Rockville, MD 20850, USA

Abbreviations: CBB, Coomassie Brilliant Blue; CPK, creatine phosphokinase; 2-D, two-dimensional; IEF, isoelectric focusing; MSN, master spot number; NP-40, Nonidet P-40; SDS, sodium dodecyl sulfate

ture and the associated shift to strong selection for growth, and how do these affect experimental outcomes? Hence the apparent advantages of *in vitro* systems, in terms of experimental manipulation, may be counterbalanced by other factors relating to 2-D data quality.

There is a second important class of reasons for exploring the use of an *in vivo* biological system such as the liver. Historically, there have been two broad approaches to the mechanistic dissection of biochemical processes in intact cellular systems: genetics (a search for informative mutants) and the use of chemical agents (drugs and chemical toxins). Both approaches help us to understand complex systems by disrupting some specific functional element and showing us the result. With the development of techniques for genetic manipulation and cloning, the genetic approach can be effectively applied either *in vitro* or *in vivo*, although the *in vitro* route is usually quicker. The chemical approach can also be applied to either sort of biological system; here, however, the bulk of consistently acquired information is in experimental animals (rats and mice). While most biologists know a short list of compounds having specific, experimentally useful effects (e.g., inhibitors of protein synthesis, ionophores, polymerase inhibitors, channel blockers, nucleotide analogs, and compounds affecting polymerization of cytoskeletal proteins), there is a much larger number of interesting chemically-induced effects, most of them characterized by toxicologists and pharmacologists in rodent systems. Just as a thorough genetic analysis would involve saturating a genome with mutations, it is possible to imagine a saturating number of drugs, the analysis of whose actions would reveal the complete biochemistry of the cell. While organized drug discovery efforts usually target specific desired effects, the nature of the process, with its dependence on screening large numbers of compounds, necessarily produces many unanticipated effects. It is therefore reasonable to suppose that the required broad range of compounds necessary to achieve "biochemical saturation" may be forthcoming: in fact, it may already exist among the hundreds of thousands of compounds that failed to qualify as drugs.

Among organs, the liver is an obvious choice for the study of chemical effects because of its well-known plasticity and responsiveness. The brain appears to be quite plastic (e.g. [7]), but it is a complicated mixture of cell types requiring skillful dissection for most experiments. The kidney, while quite responsive, also presents a potentially confounding mixture of cell types. The liver, by contrast, is made up of one predominant cell type which is easy to solubilize: the hepatocyte, representing more than 95% of its mass. Most importantly, the liver performs many homeostatic functions that require rapid modulation of gene expression. It appears that most chemical agents tested affect gene expression in the liver at some dosage (N. Leigh Anderson, unpublished observations), an interesting contrast to our earlier work with lymphocytes, for example, which seem to be much less responsive. Such results conform to the expectation that cells with a homeostatic, physiological role should be more plastic than cells differentiated for a purpose dependent on the action of a limited number of specific genes.

The liver also allows the parallels between *in vitro* and *in vivo* systems to be examined in detail. Significant progress

has been made in the development of mouse, rat and human hepatocyte culture systems, as well as in precision tissue slices. Using such an array of techniques, it is possible to assemble a matrix of mammalian systems including mouse and rat *in vivo* on one level and mouse, rat and human *in vitro* on a second level, and to compare effects between species and between systems. This approach allows us to draw informed conclusions regarding the biochemical "universality" of biological responses among the mammalia and to offer some insight into the validity of *in vitro* approaches for toxicological screening. We believe this will be necessary if *in vitro* alternatives are to achieve widespread usage in government-mandated safety testing of drugs, consumer products and industrial and agricultural chemicals.

A number of interesting studies have been published using 2-D mapping to examine effects in the rodent liver. A number of investigators have made use of the technique to screen for existing genetic variants [8-11] or induced mutations [12-14], mainly in the mouse. This work builds on the wealth of genetic information available on the mouse and its established position as a mammalian mutation-detection system. While some studies of chemical effects have been undertaken in the mouse [15-17], most have used the rat [18-23]. The examination of the cytochrome p-450 system, in particular, has been carried out almost exclusively on the rat [24, 25].

These considerations lead us to conclude that rodent liver offers the best opportunity to systematically examine an array of gene regulation systems, and ultimately to build a predictive model of large-scale mammalian gene control. The basic underlying foundation of such a project is a reliable, reproducible master 2-D pattern of liver, to which ongoing experimental results can be referred. In this paper, we report such a master pattern for the acidic and neutral proteins of rat liver (pattern F344MST3). In future, this master will be supplemented by maps of basic proteins, and analogous maps of mouse and human liver.

2 Materials and methods

2.1 Sample preparation

Liver is an ideal sample material for most biochemical studies, including 2-D analysis. A sample is taken of approximately 0.5 g of tissue from the apical end of the left lobe of the liver. Solubilization is effected as rapidly as practical; a delay of 5-15 min appears to cause no major alteration in liver protein composition if the liver pieces are kept cold (e.g., on ice) in the interim. In the solubilization process, the liver sample is weighed, placed in a glass homogenizer (e.g., 15 mL Wheaton); 8 volumes of solubilizing solution*

* The solubilizing solution is composed of 2% NP-40 (Sigma), 9 M urea (analytical grade, e.g., BDH or Bio-Rad), 0.5% dithiothreitol (DTT; Sigma) and 2% carrier ampholytes (pH 9-11 LKB; these come as a 20% stock solution, so 2% final concentration is achieved by making the final solution 10% 9-11 Ampholine by volume). A large batch of solubilizer (several hundred mL) is made and stored frozen at -80°C in aliquots sufficient to provide enough for one day's estimated sample preparation requirement. The solution is never allowed to become warmer than room temperature at any stage during preparation or thawing for use, since heating of concentrated urea solutions can produce contaminants that covalently modify proteins producing artifactual charge shifts. Once thawed, any unused solubilizer is discarded.

added (i.e., 4 mL per 0.5 g tissue) and the mixture is homogenized using first the loose- and then the tight-fit glass pestle. This takes approximately 5 strokes with the pestle and is carried out at room temperature because it would crystallize out in the cold. Once the liver sample is thoroughly homogenized in the solubilizer, it is assumed that all the proteins are denatured (by the chaotropic effect of the urea and NP-40 detergent) and the enzymes inactivated by the high pH (~9.5). Therefore these samples may be kept at room temperature until they can be centrifuged frozen as a group (within several hours of preparation). The samples are centrifuged for 6×10^4 g min (e.g., 500 000 g for 12 min using a Beckman TL-100 centrifuge). The centrifuge rotor is maintained at just below room temperature (e.g., 15–20°C), but not too cold, so as to prevent the precipitation of urea. The centrifuge of choice is a Beckman TL-100 because of the sample tube sizes available, but any ultracentrifuge accepting smallish tubes will suffice. When an appropriate centrifuge is not available near the site of sample preparation, samples can be frozen at –80°C and thawed prior to centrifugation and collection of supernatants. Each supernatant is carefully removed following centrifugation and aliquoted into at least 4 clean tubes for storage. This is done by transferring all the supernatant to one clean tube, mixing this gently (to assure homogeneous composition) and then dividing it into 4 aliquots. The aliquots are frozen immediately at –80°C. These multiple aliquots can provide insurance against a failed run or a freezer breakdown.

2. Two-dimensional electrophoresis

Sample proteins are resolved by 2-D electrophoresis using the 20 × 25 cm Iso-Dalt[®] 2-D gel system ([26–29]; produced by LSB and by Hoefer Scientific Instruments, San Francisco) operating with 20 gels per batch. All first-dimensional isoelectric focusing (IEF) gels are prepared using the same single standardized batch of carrier ampholytes BDH 4–8A in the present case, selected by LSB's batch-testing program for rat and mouse database work^{***}). A 10 µL sample of solubilized liver protein is applied to each gel, and the gels are run for 33 000 to 34 500 volt-hours using a progressively increasing voltage protocol implemented by a programmable high-voltage power supply. An 'Angelique' computer-controlled gradient-casting system (produced by LSB) is used to prepare second-dimensional sodium dodecyl sulfate (SDS) polyacrylamide gradient slab gels in which the top 5% of the gel is 11%T acrylamide, and the lower 95% of the gel varies linearly from 11% to 18%T.

This system has recently been modified so as to employ a commercially available 30.8%T acrylamide/*N,N*'-methylenebisacrylamide prepared solution (thus avoiding the handling of the solid acrylamide monomer) and three additional stock solutions: buffer (made from Sigma pre-set Tris), persulfate and *N,N,N,N*'-tetramethylethylenediamine (TEMED). Each gel is identified by a computer-printed filter paper label polymerized into the lower left corner of the gel. First-dimensional IEF tube gels are loaded

directly (as extruded) onto the slab gels without equilibration, and held in place by polyester fabric wedges ('Wedgies', produced by LSB) to avoid the use of hot agarose. Second-dimensional slab gels are run overnight, in groups of 20, in cooled DALT tanks (10°C) with buffer circulation. All run parameters, reagent source and lot information, and notations of deviation from expected results are entered by the technician responsible on a detailed, multi-page record of the experiment.

2.3 Staining

Following SDS-electrophoresis, slab gels are stained for protein using a colloidal Coomassie Blue G-250 procedure in covered plastic boxes, with 10 gels (totalling approximately 1 L of gel) per box. This procedure (based on the work of Neuhoft [30, 31]) involves fixation in 1.5 L of 50% ethanol and 2% phosphoric acid for 2 h, three 30 min washes, each in 2 L of cold tap water, and transfer to 1.5 L of 34% methanol, 17% ammonium sulfate and 2% phosphoric acid for 1 h, followed by the addition of a gram of powdered Coomassie Blue G-250 stain. Staining requires approximately 4 days to reach equilibrium intensity, whereupon gels are transferred to cool tap water and their surfaces rinsed to remove any particulate stain prior to scanning. Gels may be kept for several months in water with added sodium azide. The water washes remove ethanol that would dissolve the stain (and render the system noncolloidal, with high backgrounds). The concentrated ammonium sulfate and methanol solution is diluted by equilibration with the water volume of the gels to automatically achieve the correct final concentrations for colloidal staining. Practical advantages of this staining approach can be summarized as follows: (i) the low, flat background makes computer evaluation of small spots (max OD < 0.02) possible, especially when using laser densitometry; (ii) up to 1500 spots can be reliably detected on many gels (e.g., rat liver) at loadings low enough to preserve excellent resolution; and (iii) reproducibility appears to be very good: at least several hundred spots have coefficients of reproducibility less than 15%. This value is at least as good as previous CBB methods, and significantly better than many silver stain systems.

2.4 Positional standardization

The carbamylated rabbit muscle creatine phosphokinase (CPK) standards [32] are purchased from Pharmacia and BDH. Amino acid compositions, and numbers of residues present in proteins used for internal standardization, are taken from the Protein Identification Resource (PIR) sequence database [33].

2.5 Computer analysis

Stained slab gels are digitized in red light at 134 micron resolution, using either a Molecular Dynamics laser scanner (with pixel sampling) or an Eikonix 78/99 CCD scanner. Raw digitized gel images are archived on high-density DAT tape (or equivalent storage media) and a greyscale video-print prepared from the raw digital image as hard-copy backup of the gel image. Gels are processed using the Kepler[®] software system (produced by LSB), a commercially available workstation-based software package built on

^{***} This material (succeeding certified batches of which are available from Hoefer Scientific Instruments) has the most linear pH gradient produced by any ampholyte tested except for the Pharmacia wide range (which has an unacceptable tendency to bind high-molecular weight acidic proteins, causing them to streak).

some of the principles of the earlier TYCHO system [34-41]. Procedure PROC008 is used to yield a spotlist giving position, shape and density information for each detected spot. This procedure makes use of digital filtering, mathematical morphology techniques and digital masking to remove the background, and uses full 2-D least-squares optimization to refine the parameters of a 2-D Gaussian shape for each spot. Processing parameters and file locations are stored in a relational database, while various log files detailing operation of the automatic analysis software are archived with the reduced data. The computed resolution and level of Gaussian convergence of each gel are inspected and archived for quality control purposes.

Experiment packages are constructed using the Kepler experiment definition database to assemble groups of 2-D patterns corresponding to the experimental groups (e.g., treated and control animals). Each 2-D pattern is matched to the appropriate "master" 2-D pattern (pattern F344MST3 in the case of Fischer 344 rat liver), thereby providing linkage to the existing rodent protein 2-D databases. The software allows experiments containing hundreds of gels to be constructed and analyzed as a unit, with up to 100 gels displayed on the screen at one time for comparative purposes and multiple pages to accommodate experiments of > 1000 gels. For each treatment, proteins showing significant quantitative differences vs. appropriate controls are selected using group-wise statistical parameters (e.g., Student's *t*-test, Kepler² procedure STUDENT). Proteins satisfying various quantitative criteria (such as $P < 0.001$ difference from appropriate controls) are represented as highlighted spots onscreen or on computer-plotted protein maps and stored as spot populations (i.e., logical vectors) in a liver protein database. Quantitative data (spot parameters, statistical or other computed values) are stored as real-valued vectors in the database. Analysis of coregulation is performed using a Pierson product-moment correlation (Kepler procedure CORREL) to determine whether groups of proteins are coordinately regulated by any of the treatments. Such groups can be presented graphically on a protein map, and reported together with the statistical criteria used to assess the level of coregulation. Multivariate statistical analysis (e.g., principal components' analysis) is performed on data exported to SAS (SAS Institute).

2.6 Graphical data output

Graphical results are prepared in GKS and translated within Kepler² into output for any of a variety of devices. Linedrawing output is typically prepared as Postscript and printed on an Apple Laserwriter. Detailed maps presented here have been generated using an ultra-high-resolution Postscript-compatible Linotronic output device. Greyscale graphics are reproduced from the workstation screen using a Seikosha videoprinter. Patterns are shown in the standard orientation, with high molecular mass at the top and acidic proteins to the left.

2.7 Experiment LSBC04

In the study described here 12-week-old Charles River male F344 rats were used. Diets were prepared at LSB, based on a Purina 5755M Basal Purified Diet. Lovastatin and cholestyramine were obtained as prescription pharma-

ceuticals, ground and mixed with the diet at concentrations of 0.075% and 1%, respectively. The high cholesterol diet was Purina 5801M-A (5% cholesterol plus 1% sodium cholethiolate in the control diet). Animal work was carried out by Microbiological Associates (Bethesda, MD). Animals were acclimatized for one week on the control diet, fed test or control diets for one week, and sacrificed on day 8. Average daily doses of lovastatin and cholestyramine in appropriate groups were 37 mg/kg/day and 5 g/kg/day, respectively, based on the weight of the food consumed. Liver samples were collected and prepared for 2-D electrophoresis according to the standard liver protocol (homogenization in 8 volumes of 9 M urea, 2% NP-40, 0.5% dithiothreitol, 2% LKB pH 9-11 carrier ampholytes, followed by centrifugation for 30 min at 80 000 × *g*). Kidney, brain and plasma samples were frozen. Gels were run as described above, and the data was analyzed using the Kepler² system. Gels were scaled, to remove the effect of differences in protein loading, by setting the summed abundances of a large number of matched spots equal for each gel (linear scaling).

3 Results and discussion

3.1 The rat liver protein 2-D map

F344MST3 is a standard 2-D pattern of rat liver proteins, based on the Fischer 344 strain. This pattern was initiated from a single 2-D gel and extensively edited in an experiment comparing it to a range of protein loads, so as to include both small spots and well-resolved representations of high-abundance spots. More than 700 rat liver 2-D patterns have been matched to F344MST3 in a series of drug effects and protein characterization experiments, and numerous new spots (induced by specific drugs, for instance) have been added as a result. A modified version including additional spots present in the Sprague-Dawley outbred rat has also been developed (data not shown). Figure 1 shows a greyscale representation and Fig. 2 a schematic plot of the master pattern. More than 1200 spots are included, most of which are visible on typical gels loaded with 10 µL of solubilized liver protein prepared by the standard method and stained with colloidal Coomassie Blue. Master spot numbers (MSN's) have been assigned to all proteins, and appear in the following figures, each showing one quadrant of the pattern. Figure 3 shows the upper left (acidic, high molecular mass) quadrant, Fig. 4 the upper right (basic, high molecular mass) quadrant, Fig. 5 the lower left (acidic, low molecular mass) quadrant, and Fig. 6 the lower right (basic, low molecular mass) quadrant. The quadrants overlap as an aid to moving between them. The gel position (in 100 micron units), isoelectric point (relative to the CPK internal *pI* standards) and SDS molecular mass (from the calibration curve in Fig. 8) are listed for each spot (Table 1). Because of the precision of the CPK-*pI* values, these parameters can be used to relate spot locations between gel systems more reliably than using *pI* measurements expressed as pH. A major objective of current studies is the identification of all major spots corresponding to known liver proteins, as well as rigorous definitions of subcellular organelle contents. Of particular interest to us is the parallel development of identifications in the rat and mouse liver maps, allowing detailed comparisons of gene expression effects in the two systems. The results of these studies will be presented systematically in a later edition of this database.

We include here a useful series of 22 orienting identifications as an aid to other users of the rat liver pattern (Table 1).

2 Carbamylated charge standards, computed pI's and molecular mass standardization

We have previously shown that the use of a system of close-spaced internal pI markers (made by carbamylating a basic protein) offers an accurate and workable solution to the problem of assigning positions in the pI dimension [32]. The same system, based on 36 protein species made by carbamylating rabbit muscle CPK, has been used here to assign pI's to most rat liver acidic and neutral proteins. The standards were coelectrophoresed with total liver proteins, and the standard spots added to a special version of the master pattern F344MST3. The gel X-coordinates of all liver protein spots lying within the CPK charge train were then transformed into CPK pI positions by interpolation between the positions of immediately adjacent standards (Table 1) using a Kepler² vector procedure.

It has proven possible to compute fairly accurate pI values for many proteins from the amino acid composition [42]. We have attempted here to test a further elaboration of this approach, in which we computed pI's for the CPK standards themselves, based on our knowledge of the rabbit muscle CPK sequence and the fact that adjacent members of the charge train typically differ by blockage of one additional lysine residue (Table 3). We compared these values to similar computed pI's for an additional set of carbamylated standards made from human hemoglobin beta chains and a series of rat liver and human plasma proteins of known position and sequence (Fig. 7, Table 4). The result demonstrates good concordance between these systems. Two proteins show significant deviations: liver fatty-acid binding protein (FABP; #1 in Table 4) and protein disulphide isomerase (#20 in the table). The FABP spot present on F344MST3 may represent a charge-modified version of a more basic parent spot closer to the expected pI, not resolved in the IEF/SDS gel. Of particular importance is the fact that, by comparing computed pI's of sequenced but unlocated proteins with the CPK pI's, we can assign a probable gel location without making any assumptions regarding the actual gel pH gradient. This offers a useful shortcut, given the vagaries of pH measurement on small diameter IEF gels. We have used this approach to compute the CPK pI's of all rat and mouse proteins in the PIR sequence database, as an aid to protein identification (data not shown).

In order to standardize SDS molecular weight (SDS-MW), we have used a standard curve fitted to a series of identified proteins (Fig. 8). Rather than using molecular mass *per se*, we have elected to use the number of amino acids in the polypeptide chain, as perhaps a better indication of the length of the SDS-coated rod that is sieved by the second dimension slab. The resulting values were multiplied by 112 (the weighted average mass of amino acids in sequenced proteins) to give predicted molecular masses. Because we use gradient slabs, we have not constrained the fit curve to conform to any predetermined model; rather we tried many equations and selected the best using the program "Tablecurve" on a PC. The equation chosen was $y = a + bx + cx^2$, where y is the number of residues, x is the gel

Y coordinate, a is 511.83, b is -0.2731 and c is 33183801. The resulting fit appears to be fairly good over a broad range of molecular mass.

3.3 An example of rat liver gene regulation: Cholesterol metabolism

Experiment LSBC04 was designed as a small-scale test of the regulation of cholesterol metabolism *in vivo* by three agents included in the diet: lovastatin (Mevacor², an inhibitor of HMG-CoA reductase); cholestyramine (a bile acid sequestrant that has the effect of removing cholesterol from the gut-liver recirculation); and cholesterol itself. The first two agents should lower available cholesterol and the third should raise it, allowing manipulation of relevant gene expression control systems in both directions. Such an experiment offers an interesting test of the 2-D mapping system since most of the pathway enzymes are present in low abundance, many are membrane-bound and difficult to solubilize, and the pathway itself is complex. Approximately 1000 proteins were separated and detected in liver homogenates. Twenty-one proteins were found to be affected by at least one treatment, and these could be divided into several coregulated groups.

3.3.1 MSN 413 (putative cytosolic HMG-CoA synthase) and sets of spots regulated coordinately or inversely

One group of spots (including a spot assigned to the cytosolic HMG-CoA synthase, MSN 413) showed the expected increase in abundance with lovastatin or cholestyramine, the synergistic further increase with lovastatin and cholestyramine, and a dramatic decrease with the high cholesterol diet. Spot number 413 is the most strongly regulated protein in the present experiment, showing a 5- to 10-fold induction after a 1 week treatment with 0.075% lovastatin and 1% cholestyramine in the diet (Figs. 9 and 10). Its expression follows precisely the expectation for an enzyme whose abundance is controlled by the cholesterol level; it is progressively increased from the control levels by cholestyramine, lovastatin and lovastatin plus cholestyramine, and it sinks below the threshold of detection in animals fed the high cholesterol diet. This spot has been tentatively identified as the cytosolic HMG-CoA synthase, based on a reaction with an antiserum to that protein provided by Dr. Michael Greenspan at Merck Sharp & Dohme Research Laboratories. This enzyme lies immediately before HMG-CoA reductase in the liver cholesterol biosynthesis pathway, and is known to be co-regulated with it. Spot 413 has an SDS molecular weight of about 54 000 and a CPK pI of -11.4, in reasonably close agreement with a molecular weight of 57300 and a CPK pI of -15.7 computed from the known sequence of the hamster enzyme [43].

Using a classical product-moment correlation test (Kepler procedure CORREL), a series of five additional spots was found to be coregulated with 413. The level of correlation was exceedingly high (> 95%). Two of these, 1250 and 933, are at similar molecular weights and approximately one charge more acidic than 413 (Fig. 9), indicating that they may be covalently modified forms of the 413 polypeptide. This suspicion is strengthened by the observation that both spots are also stained by the antibody to cytosolic HMG-CoA synthase. The remaining three correlated spots appear

to comprise an additional related pair (1253 and 1001) of around 40 kDa and a single spot (1119) of around 28 kDa. Because these two presumed proteins are present at substantially lower abundances than 413, and because the cytosolic HMG-CoA synthase is reported to consist of only one type of polypeptide, they are likely to represent other, very tightly coregulated enzymes. A second group of six spots was selected based on a regulatory pattern close to the inverse of that for spot 413 (MSN's 34, 79, 178, 182, 204, 347; data not shown). For these proteins, the lowest level of expression occurs with exposure to lovastatin plus cholestyramine and the highest level upon exposure to the high-cholesterol diet. Spots 182 and 79 are highly correlated and lie about one charge apart at the same molecular weight; they may thus be isoforms of a single protein. The other four spots probably represent additional enzymes or subunits.

3.3.2 MSN 235 and coregulated spots

A third group of five spots, mainly comprised of mitochondrial proteins including putative mitochondrial HMG-CoA synthase spots, showed a modest induction by lovastatin alone, but little or no effect with any of the other treatments (including the combination of lovastatin and cholestyramine; Fig. 12). This result is intriguing because lovastatin was expected to affect only the regulation of enzymes of cholesterol synthesis, which is entirely extra-mitochondrial. Three of the spots (235, 134, 144) form a closely-packed triad at approximately 30 kDa, and are likely to represent isoforms of one protein. All three spots are stained by an antibody to the mitochondrial form of HMG-CoA synthase obtained from Dr. Greenspan. Subcellular fractionation indicates a mitochondrial location. The other two spots (633 at about 38 kDa and 724 at about 69 kDa) are each present at lower abundance than the members of the triad.

3.3.3 An example of an anti-synergistic effect

A sixth spot (367) shows strong induction by lovastatin (two- to threefold), and about half as much induction with lovastatin plus cholestyramine, but without sharing the animal-animal heterogeneity pattern of the 235-set (Fig. 13). This protein is also mitochondrial, and represents the clearest example of an anti-synergistic effect of lovastatin and cholestyramine. The existence of such an effect demonstrates that lovastatin and cholestyramine do not act exclusively through the same regulatory pathway.

3.3.4 Complexity of the cholesterol synthesis pathway

Taken together, these results suggest that treatment with lovastatin alone can affect both cytosolic and mitochondrial pathways using HMG-CoA, while cholestyramine, on the other hand, either alone or in combination with lovastatin, produces a strong effect on the putative cytosolic pathway, but little or no effect on the putative mitochondrial pathway. An explanation for this difference may lie in lovastatin's effect on levels of HMG-CoA and related precursor compounds that are exchanged between the cytosol and the mitochondrion, whereas cholestyramine should affect only the cytosolic pathways directly controlled by cholesterol and bile acid levels. It remains to be explained why some

proteins of the putative mitochondrial pathway are so much more variable in their expression in all groups. An examination of all the coregulated groups suggests that quantitative statistical techniques can extract a wealth of interesting information from large sets of reproducible gels. The abundance of spots in the 413 coregulation group, for example, shows an amazing level of concordance in their relative expression among the five individuals of the lovastatin and cholestyramine treatment group. This effect is not due to differences in total protein loading, since they have already been removed by scaling, and since proteins with quite different regulation patterns can be demonstrated (e.g., Fig. 13). Such effects raise the possibility that many gene coregulation sets may be revealed through the study of a sufficiently large population of control animals (i.e., without any experimental manipulation). This approach, exploiting natural biological variation in protein expression instead of drug effects, offers an important incentive for the construction of a large library of control animal patterns.

4 Conclusions

Because of the widespread use of rat liver in both basic biochemistry and in toxicology, there is a long-term need for a comprehensive database of liver proteins. The rat liver master pattern presented here has proven to be an accurate representation of this system, having been matched to more than 700 gels to date. As the number of proteins identified and the number of compounds tested for gene expression effects grows, we expect this database to contribute valuable insights into gene regulation. Its practical utility in several areas of mechanistic toxicology is already being demonstrated.

Received September 11, 1991

5 References

- [1] O'Farrell, P. J. *Biol. Chem.* 1975, 250, 4007-4021.
- [2] Klose, J. *Humangenetik* 1975, 26, 231-243.
- [3] Scheele, G. A. *J. Biol. Chem.* 1975, 250, 5375-5385.
- [4] Iborra, G. and Buhler, J. M. *Anal. Biochem.* 1976, 74, 503-511.
- [5] Anderson, N. G. and Anderson, N. L. *Behring. Inst. Mitt.* 1979, 63, 169-210.
- [6] Anderson, N. G. and Anderson, N. L. *Clin. Chem.* 1982, 28, 739-748.
- [7] Heydorn, W. E., Creed, G. J. and Jacobowitz, D. M. *J. Pharmacol. Exp. Therap.* 1984, 229, 622-628.
- [8] Anderson, N. L., Nance, S. L., Tollaksen, S. L., Giere, F. A. and Anderson, N. G. *Electrophoresis* 1985, 6, 592-599.
- [9] Racine, R. R. and Langley, C. H. *Biochem. Genet.* 1980, 18, 185-197.
- [10] Klose, J. *Mol. Evol.* 1982, 18, 315-328.
- [11] Neel, J. V., Baier, L., Hanash, S. and Erickson, R. P. *J. Hered.* 1985, 76, 314-320.
- [12] Marshall, R. R., Raj, A. S., Grant, F. J. and Heddle, J. A. *Can. J. Genet. Cytol.* 1983, 25, 457-446.
- [13] Taylor, J., Anderson, N. L., Anderson, N. G., Gemmell, A., Giometti, C. S., Nance, S. L. and Tollaksen, S. L., in: Dunn, M. J. (Ed.), *Electrophoresis '86*, Verlag Chemie, Weinheim 1986, pp. 583-587.
- [14] Giometti, C. S., Gemmell, M. A., Nance, S. L., Tollaksen, S. L. and Taylor, J. *J. Biol. Chem.* 1987, 262, 12764-12767.
- [15] Anderson, N. L., Giere, F. A., Nance, S. L., Gemmell, M. A., Tollaksen, S. L. and Anderson, N. G., in: Galteau, M.-M. and Siest, G. (Eds.), *Progrès Récents en Electrophorèse Bidimensionnelle*, Presses Universitaires de Nancy, Nancy 1986, pp. 253-260.
- [16] Anderson, N. L., Swanson, M., Giere, F. A., Tollaksen, S., Gemmell, A., Nance, S. L. and Anderson, N. G. *Electrophoresis* 1986, 7, 44-48.

- Anderson, N. L., Giere, F. A., Nance, S. L., Gemmell, M. A., Tollakson, S. L. and Anderson, N. G., *Fundam. Appl. Toxicol.* 1987, 8, 39-50.
- Anderson, N. L., in: *New Horizons in Toxicology*, Eli Lilly Symposium, 1991, in press.
- Antoine, B., Rahimi-Pour, A., Sieck, G., Magdalou, J. and Galleau, M. M., *Cell. Biochem. Funct.* 1987, 5, 217-231.
- Elliott, B. M., Ramasamy, R., Stonard, N. D. and Spragg, S. P., *Biochim. Biophys. Acta* 1986, 870, 135-140.
- Huber, B. E., Heilman, C. A., Wirth, P. J., Miller, M. J. and Thorgeirsson, S. S., *Hepatology* 1986, 6, 206-219.
- Wirth, P. J. and Vesterberg, O., *Electrophoresis* 1988, 9, 47-53.
- Witzmann, F. A. and Parker, D. N., *Toxicol. Lett.* 1991, 57, 29-36.
- Rampersaud, A., Waxman, D. J., Ryan, D. E., Levin, W. and Walz, F. G., Jr., *Arch. Biochem. Biophys.* 1985, 242, 174-183.
- Vlasuk, G. P. and Walz, F. G., Jr., *Anal. Biochem.* 1980, 105, 112-120.
- Anderson, N. G. and Anderson, N. L., *Anal. Biochem.* 1978, 85, 331-340.
- Anderson, N. L. and Anderson, N. G., *Anal. Biochem.* 1978, 85, 341-354.
- Anderson, L., Hofmann, J.-P., Anderson, E., Walker, B. and Anderson, N. G., in: Endler, A. T. and Hazashi, S. (Eds.), *Two-Dimensional Electrophoresis*, VCH Verlagsgesellschaft, Weinheim 1989, pp. 288-297.
- Anderson, L., *Two-Dimensional Electrophoresis: Operation of the ISO-DALT[®] System*, Large Scale Biology Press, Washington, DC 1988, ISBN 0-945532-00-8, 170pp.
- Neuhoff, V., Stamm, R. and Eibl, H., *Electrophoresis* 1985, 6, 427-448.
- [31] Neuhoff, V., Arold, N., Taube, D. and Ehrhardt, W., *Electrophoresis* 1988, 9, 255-262.
- [32] Anderson, N. L. and Hickman, B. J., *Anal. Biochem.* 1979, 93, 312-320.
- [33] Sidman, K. E., George, D. E., Barker, W. C. and Hunt, L. T., *Nucl. Acids Res.* 1988, 16, 1869-1871.
- [34] Taylor, J., Anderson, N. L., Coulter, B. P., Scandora, A. E. and Anderson, N. G., in: Radola, B. J. (Ed.), *Electrophoresis '79*, de Gruyter, Berlin 1980, pp. 329-339.
- [35] Taylor, J., Anderson, N. L. and Anderson, N. G., in: Allen, R. C. and Arnaud, P. (Eds.), *Electrophoresis '81*, de Gruyter, Berlin 1981, pp. 383-400.
- [36] Anderson, N. L., Taylor, J., Scandora, A. E., Coulter, B. P. and Anderson, N. G., *Clin. Chem.* 1981, 27, 1807-1820.
- [37] Taylor, J., Anderson, N. L., Scandora, A. E., Jr., Willard, K. E. and Anderson, N. G., *Clin. Chem.* 1982, 28, 861-866.
- [38] Taylor, J., Anderson, N. L. and Anderson, N. G., *Electrophoresis* 1983, 4, 338-345.
- [39] Anderson, N. L. and Taylor, J., in: *Proceedings of the Fourth Annual Conference and Exposition of the National Computer Graphics Association*, Chicago, June 26-30, 1983, pp. 69-76.
- [40] Anderson, N. L., Hofmann, J.-P., Gemmell, A. and Taylor, J., *Clin. Chem.* 1984, 30, 2031-2036.
- [41] Anderson, L., in: Schafer-Nielsen, C. (Ed.), *Electrophoresis '86*, VCH Verlagsgesellschaft, Weinheim 1988, pp. 313-321.
- [42] Neidhardt, F. C., Appleby, D. A., Sankar, P., Hutton, M. E. and Phillips, T. A., *Electrophoresis* 1989, 10, 116-121.
- [43] Gil, G., Goldstein, J. L., Slaughter, C. A. and Brown, M. S., *J. Biol. Chem.* 1986, 261, 3710-3716.

6 Addendum 1: Figures 1-13

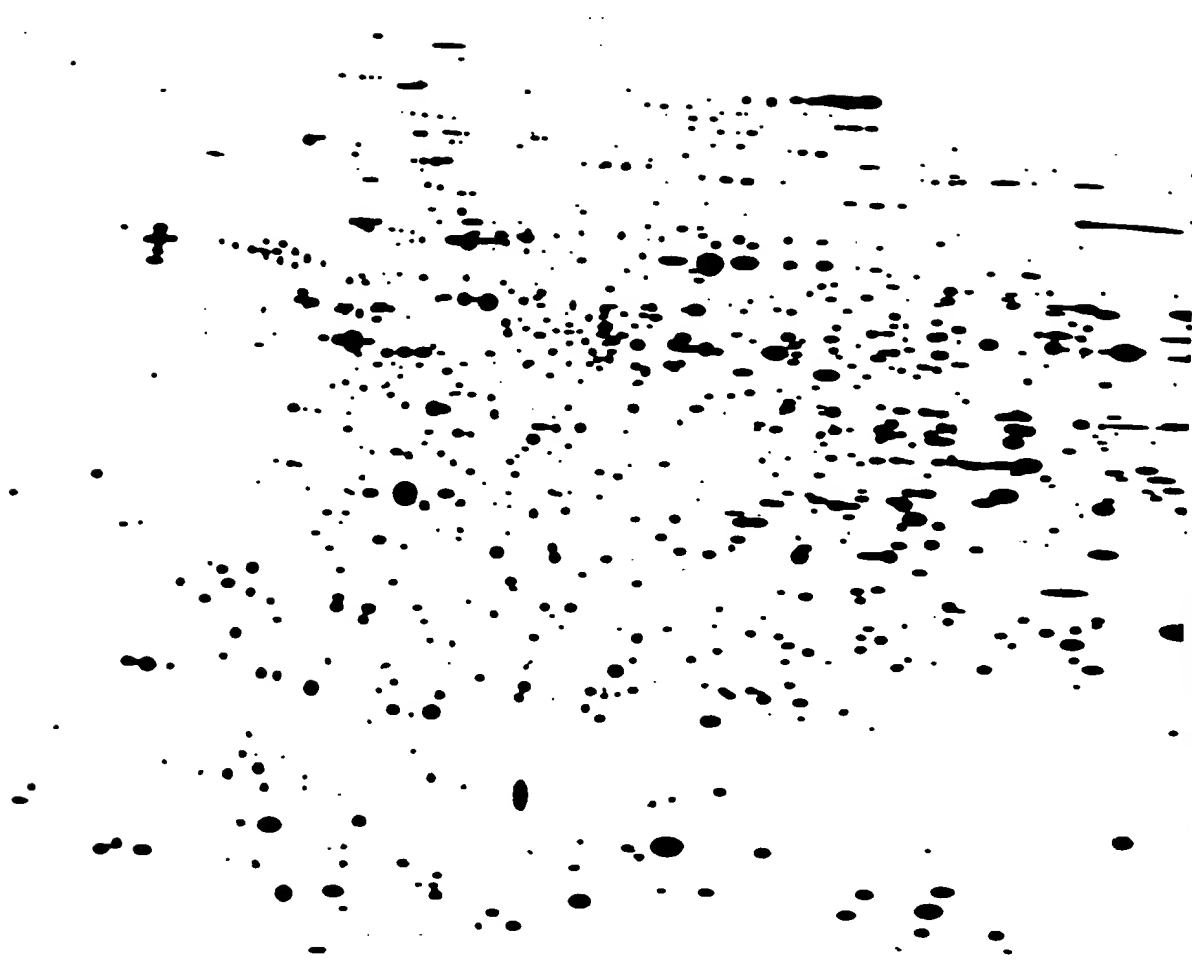


Figure 1. Synthetic representation of the standard rat liver 2-D master pattern, rendered as a greyscale image using a videoprinter.

Fig. 2. Schem
matic.

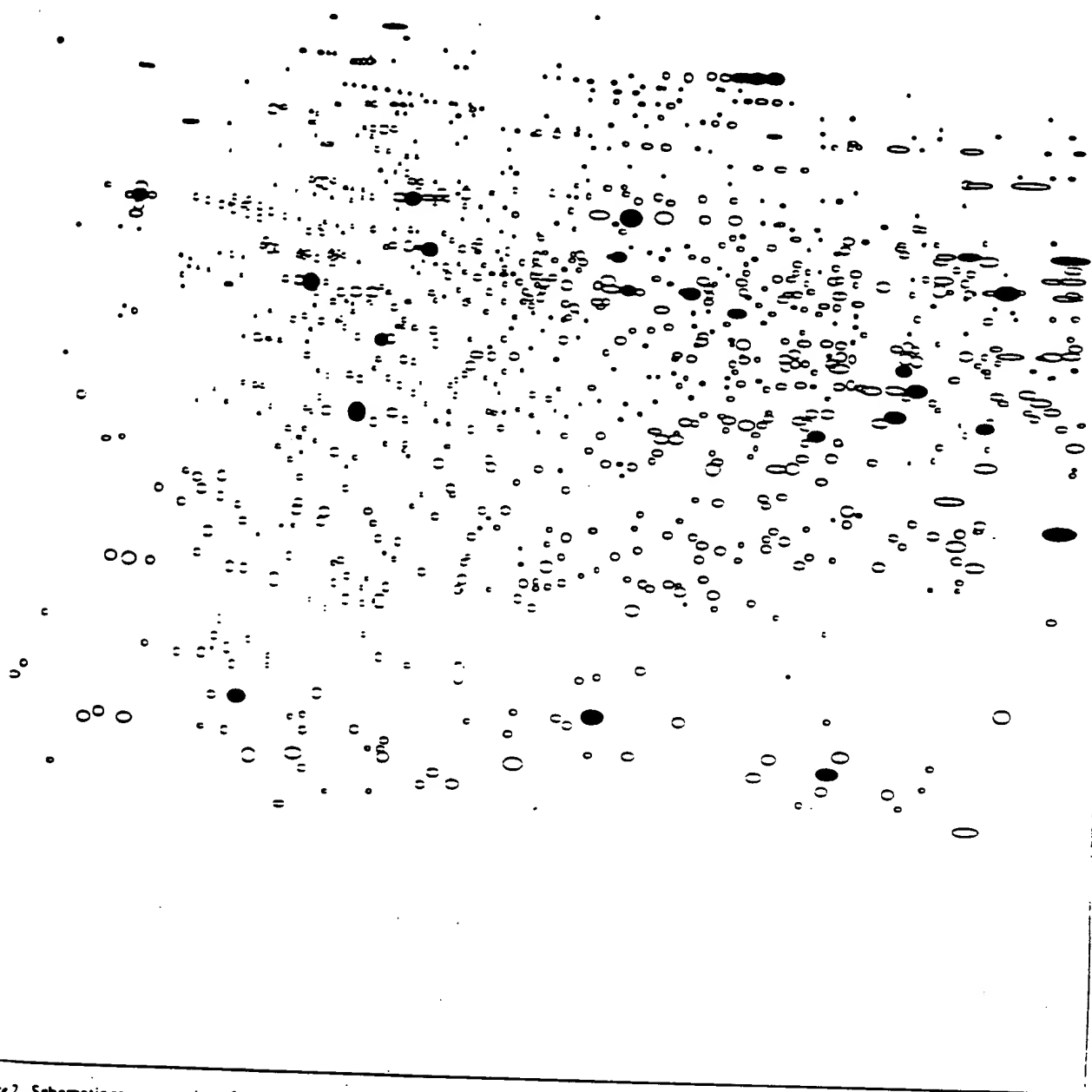


Figure 2. Schematic representation of the master pattern (the same as Fig. 1), useful as an aid in relating specific areas of Fig. 1 and the following detailed prints.

1

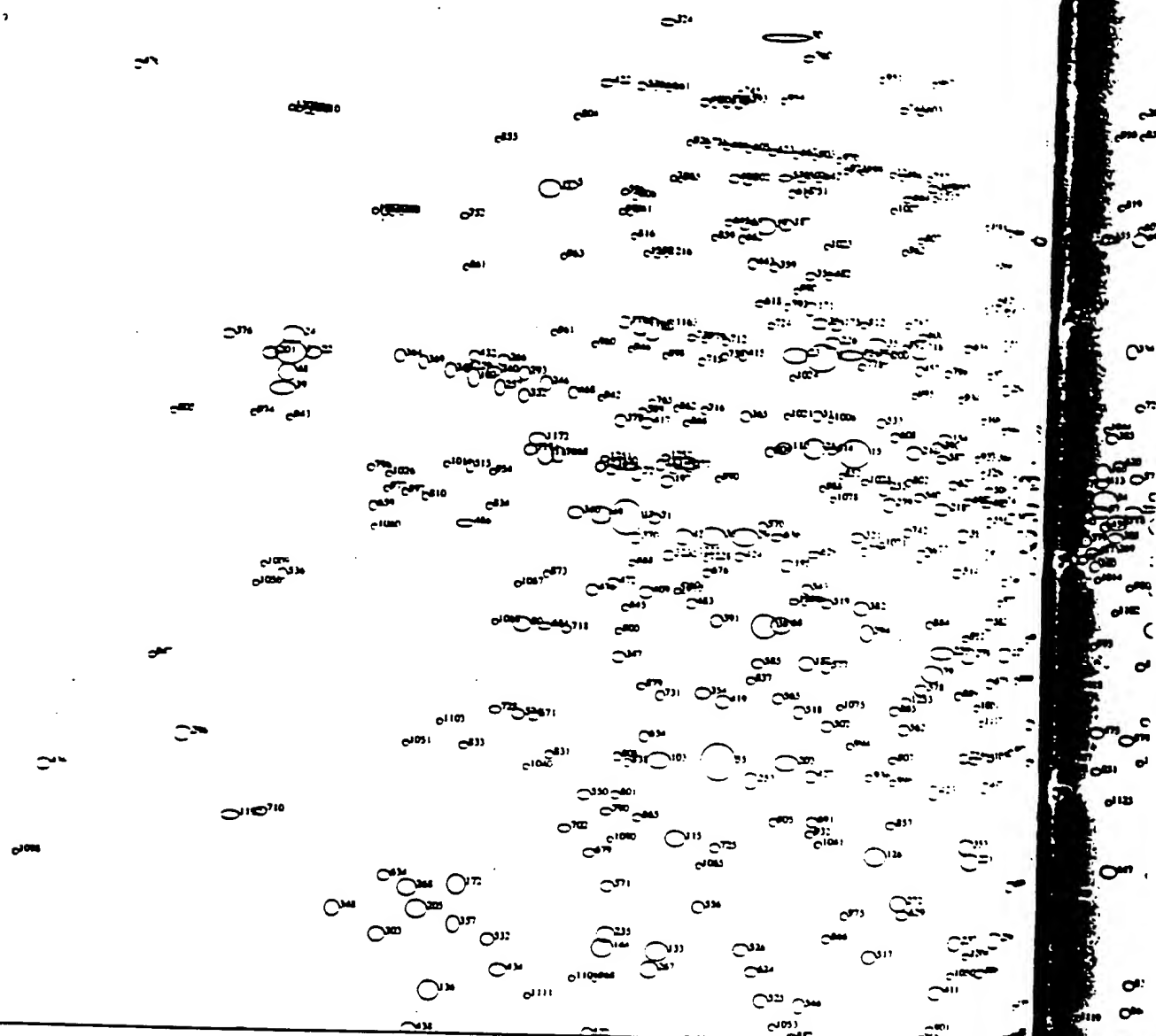


Figure 3. Upper left (high molecular weight, acidic) quadrant (#1) of the rat liver map, showing spot numbers.

4. Up1

0100

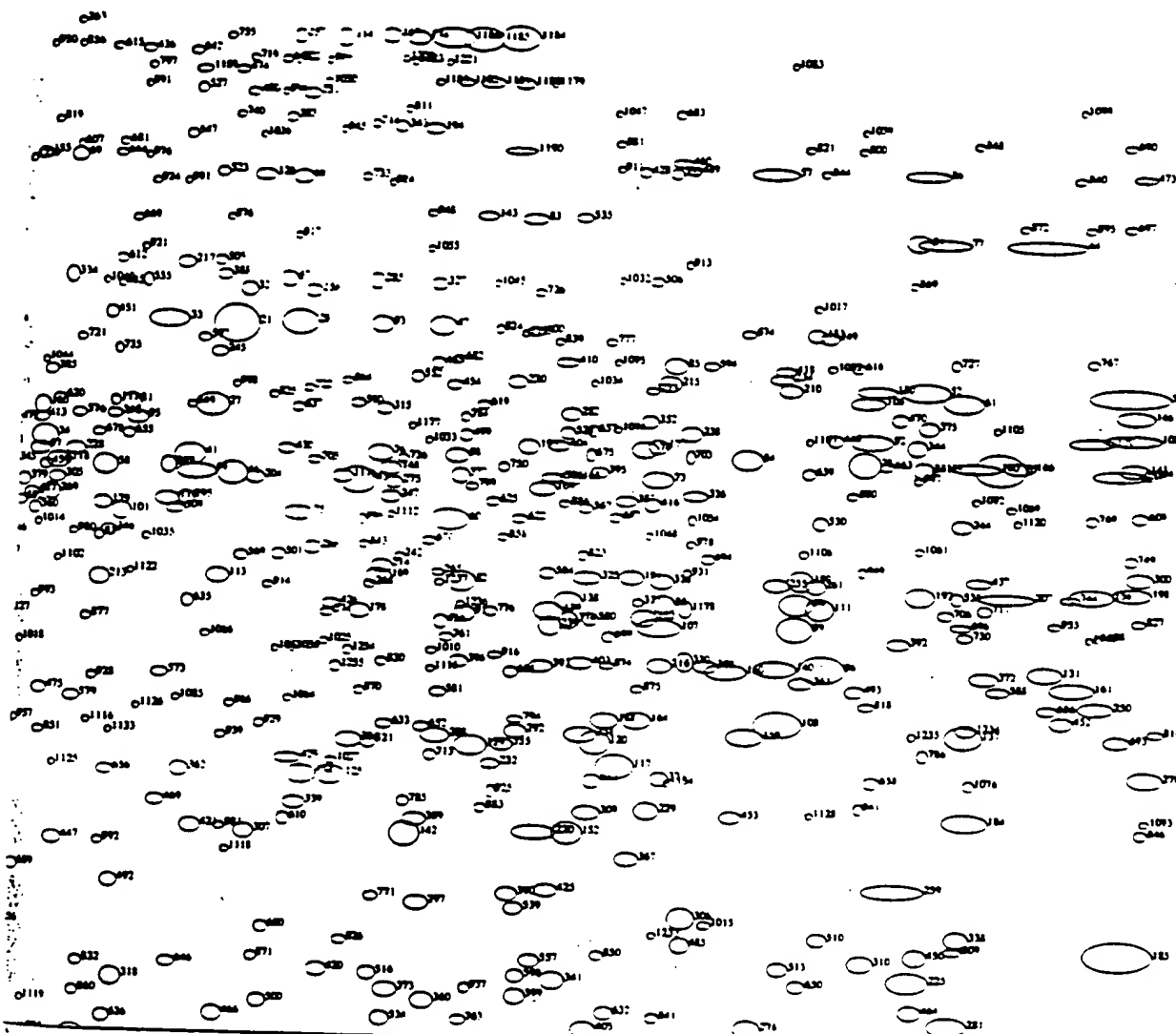
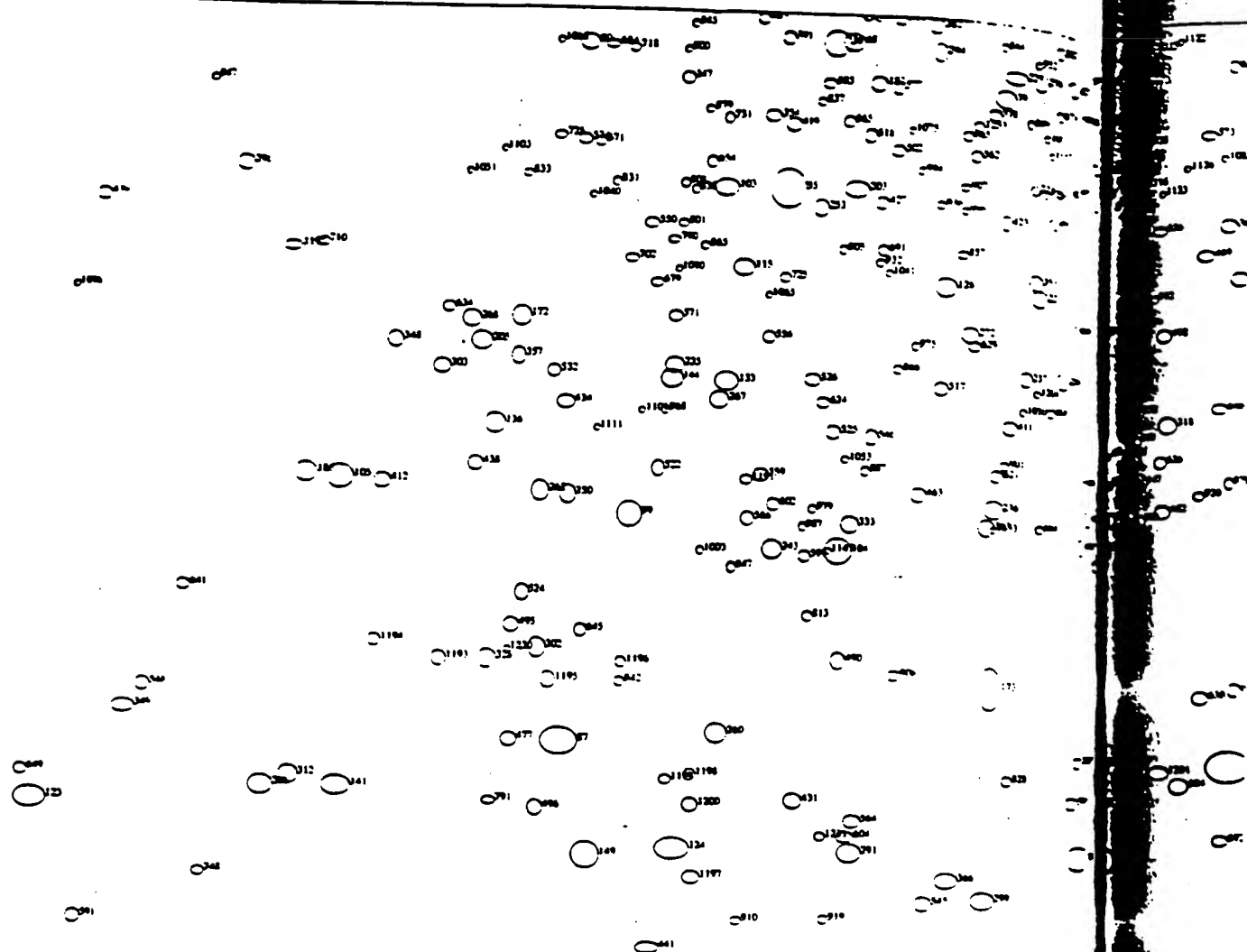


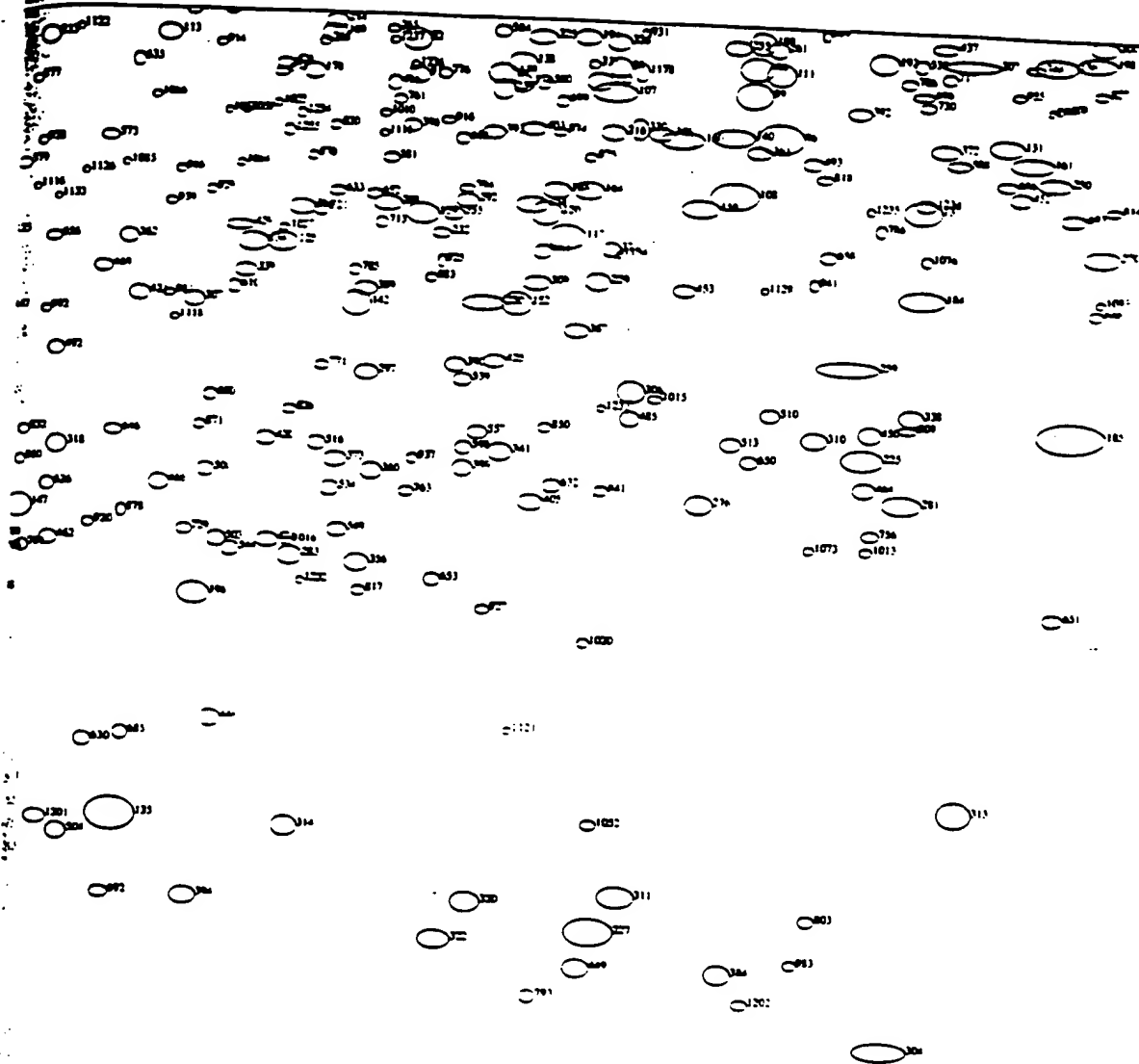
Figure 4. Upper right (high molecular weight, basic) quadrant (#2) of the rat liver map, showing spot numbers.



3

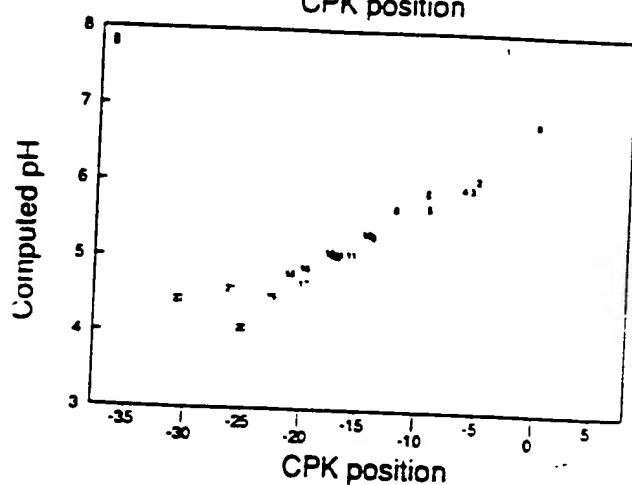
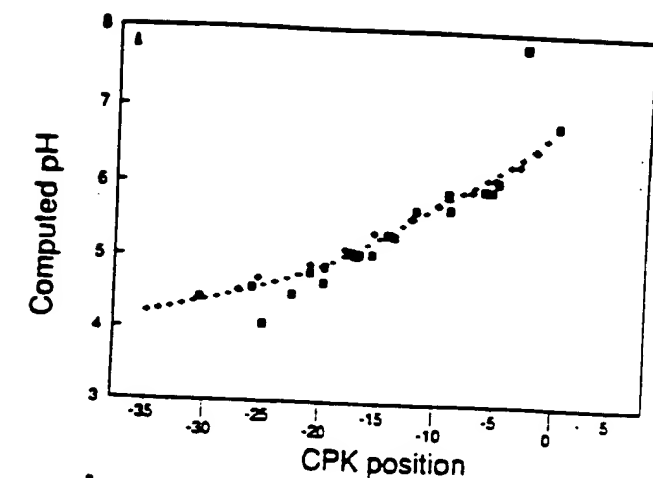
Figure 5. Lower left (low molecular weight, acidic) quadrant (#3) of the rat liver map, showing spot numbers.

6. Lower r

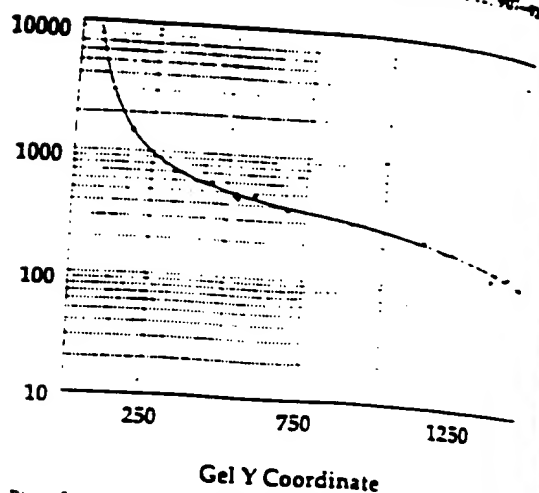


4

Figure 6. Lower right (low molecular weight, basic) quadrant (#4) of the rat liver map, showing spot numbers.



Number of Residues



Gel Y Coordinate

Figure 6. Plot of number of amino acids versus gel Y-position, with fitted curve used to predict molecular mass of unidentified proteins

Figure 7. (a) Plot of computed isoelectric point versus gel X-position for two sets of carbamylated standard proteins (rabbit muscle CPK [•] and human hemoglobin β chain, filled diamonds) and several other proteins (shaded squares). (b) The identities of the various proteins represented by the squares are indicated by the numbers in corresponding positions on (a); these refer to Table 4.

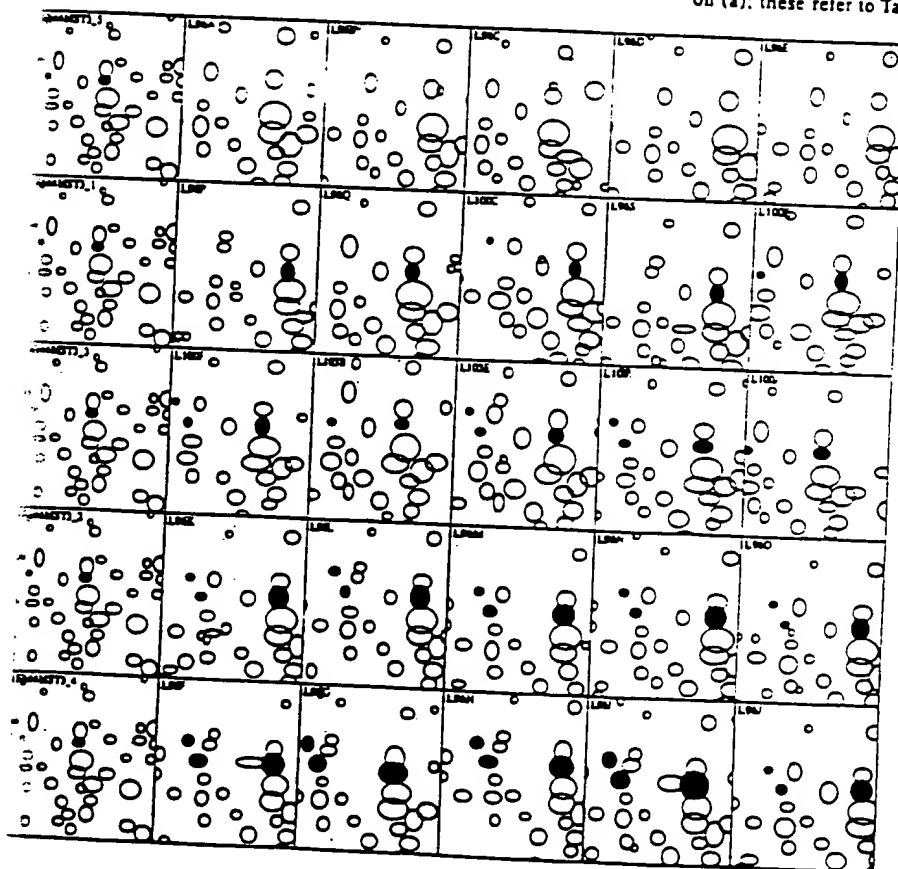


Figure 9. Montage showing effects in the region of MSN:413. The montage shows a small window into one portion of the 2-D pattern, one row of windows for each experimental group, and one panel for each gel in the experiment. The left-most pattern in each row is a group-specific copy of the master pattern followed by the patterns for the five individual rats in the group. The highlighted protein spots (filled circles) are spot 413 (on the right of each panel; identified as cytosolic HMG-CoA synthase) and two modified forms of it (1250 and 933). From the top, the rows (experimental groups) are: high cholesterol, controls, cholestyramine, lovastatin, and lovastatin plus cholestyramine.

Regulation of Rat Liver 413

(Putative Cytosolic HMG-CoA Synthase, 53kd)
Test Compounds in Diet

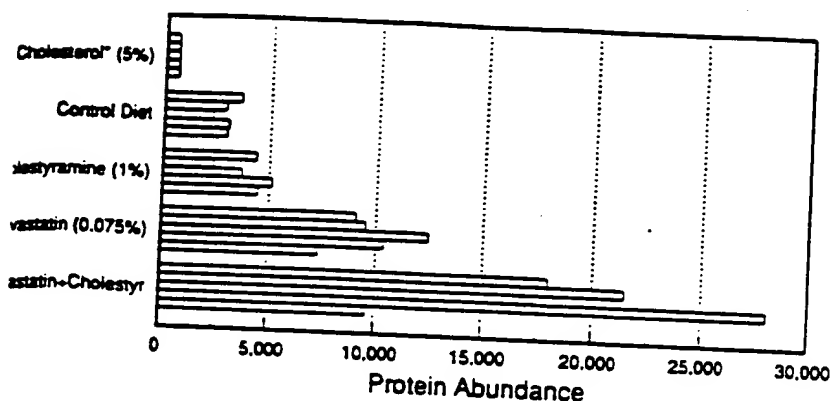


Figure 10. Bargraph showing the quantitative effects of various treatments on the abundance of MSN:413 (cytosolic HMG-CoA synthase) in the gels of Fig. 9.

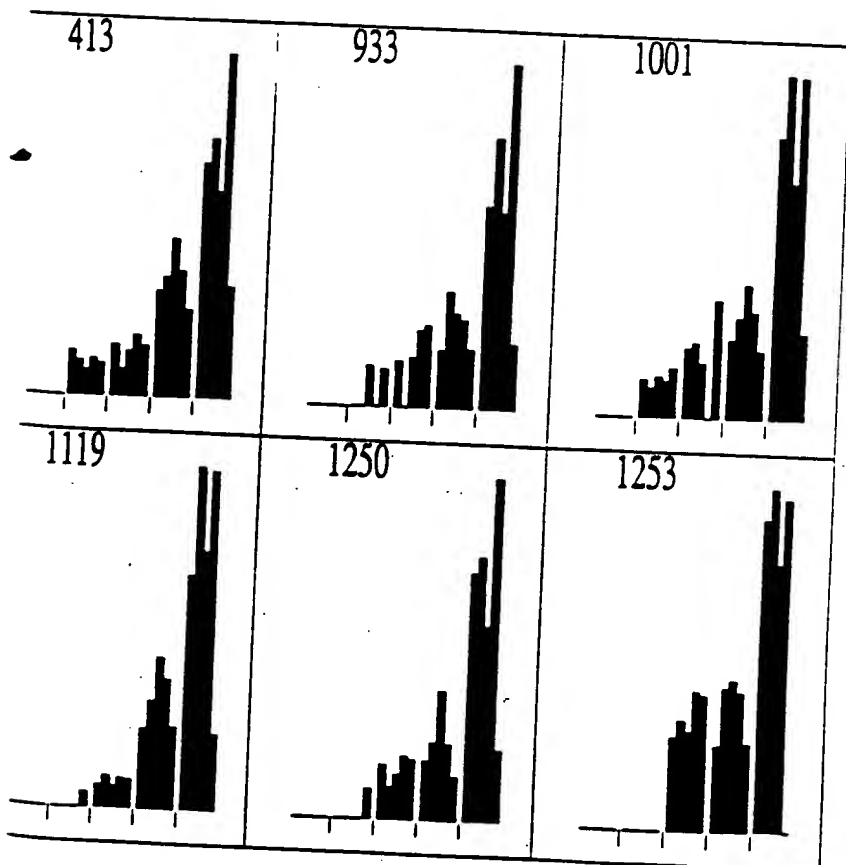


Figure 11. Bargraphs of a series of six coregulated spots including MSN:413. In the bargraphs, the abundances of the appropriate spot (master spot number shown at the top of the panel) in each animal are shown. The five five-animal groups are in the order (left to right): high cholesterol, controls, cholestyramine, lovastatin, and lovastatin plus cholestyramine. Each bar within a group represents one experimental animal liver (one 2-D gel). Note the correlated expression of the 6 spots, especially in the two far right (most strongly induced) groups.

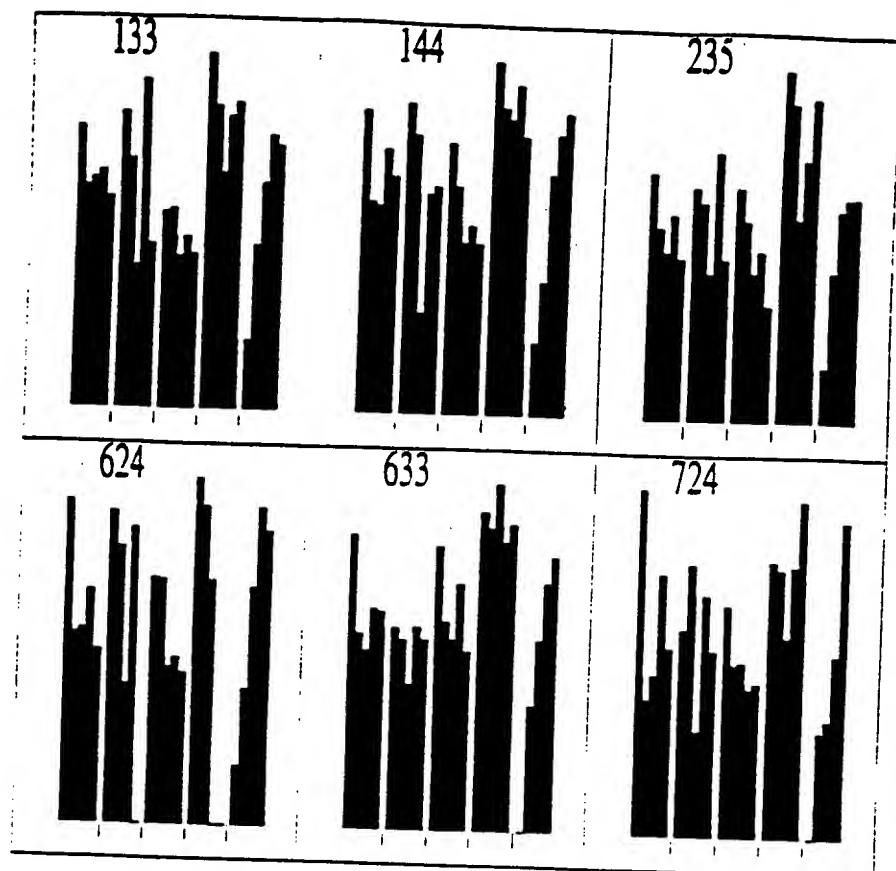


Figure 12. Data on a second coregulated group of spots, presented as in Fig. 11. The fourth experimental group (lovastatin) shows a modest induction, while the fifth group (lovastatin plus cholestyramine) does not.

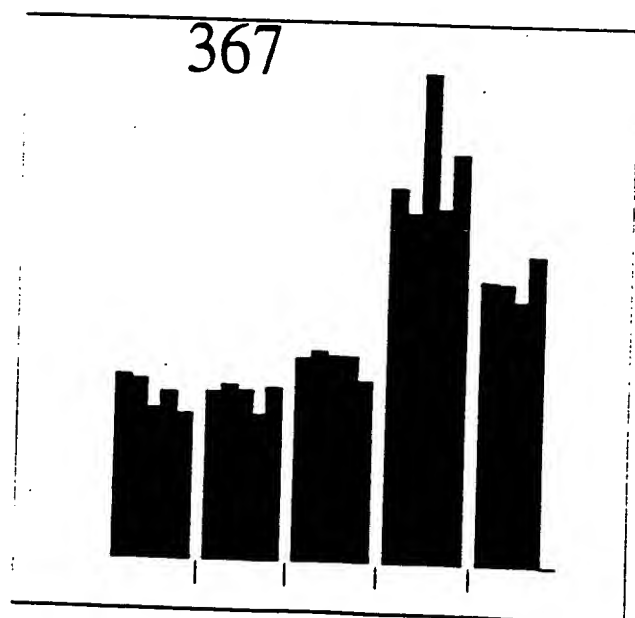


Figure 13. Data on spot MSN:367, presented as in Fig. 11. This protein shows unambiguously the anti-synergistic effect of lovastatin and cholestyramine (fifth group) as compared to lovastatin (fourth group). This response contrasts strongly with the regulation pattern seen in Fig. 11.

Appendix

Mass

x

311

508

812

846

888

908

928

948

968

988

1008

1028

1048

1068

1088

1108

1128

1148

1168

1188

1208

1228

1248

1268

1288

1308

1328

1348

1368

1388

1408

1428

1448

1468

1488

1508

1528

1548

1568

1588

1608

1628

1648

1668

1688

1708

1728

1748

1768

1788

1808

1828

1848

1868

1888

1908

1928

1948

1968

1988

2008

2028

2048

2068

2088

2108

2128

2148

2168

2188

2208

after table of;
indicated moic

Fig. 1. Master table of proteins in the rat liver database¹¹

MSN	X	Y	CPK _{el}	SDSMW	MSN	X	Y	CPK _{el}	SDSMW	MSN	X	Y	CPK _{el}	SDSMW
3	311	434	<-35.0	63,800	95	1119	536	-9.9	53,800	174	1364	183	-6.7	162,800
5	568	263	-24.3	102,900	96	1731	756	-2.0	40,700	175	825	393	-15.7	69,300
8	812	426	-16.0	64,800	97	1033	566	-11.4	51,600	177	1582	553	-3.6	52,600
11	549	268	-25.2	101,000	98	1406	565	-6.1	51,700	178	1321	710	-7.2	43,000
15	845	520	-15.3	55,200	99	578	1149	-23.8	25,000	179	1089	615	-10.4	48,300
17	629	569	-21.6	50,000	100	2004	538	>0.0	53,700	180	1866	567	-0.5	51,600
18	806	414	-14.0	86,300	101	1106	623	-10.1	47,900	181	411	295	-32.1	91,200
19	755	298	-17.5	90,200	102	482	455	-28.5	61,300	182	804	730	-16.2	42,000
20	649	403	-20.9	67,900	103	665	830	-20.2	37,300	184	1860	896	-0.6	34,500
21	1204	448	-8.7	62,100	104	773	1182	-17.0	23,800	185	1997	1017	>0.0	29,800
22	332	434	<-35.0	63,800	105	312	1117	<-35.0	26,100	186	279	1113	<-35.0	26,300
23	787	424	-16.6	65,000	106	1769	509	-1.5	56,100	187	773	296	-17.0	90,800
24	313	417	<-35.0	66,000	107	1585	720	-3.6	42,500	188	1538	807	-4.2	38,400
25	807	516	-16.1	55,500	108	1692	807	-2.4	38,300	191	1560	674	-3.9	44,900
27	1184	524	-9.0	54,900	109	1482	563	-4.8	49,700	192	1818	687	-0.9	44,200
28	1263	446	-8.0	62,400	110	778	516	-16.9	55,500	193	1469	555	-5.0	52,400
29	743	605	-17.8	49,000	111	1728	700	-2.0	43,500	194	1380	266	-6.4	101,600
30	768	112	-17.2	348,600	113	1191	680	-8.9	44,500	195	784	632	-16.7	47,300
32	1216	417	-8.6	66,000	114	1296	185	-7.5	160,800	196	1227	1185	-8.4	23,700
33	1145	445	-9.5	62,500	115	682	907	-19.6	34,100	197	667	553	-20.1	52,600
34	1037	555	-11.3	52,400	116	1146	610	-9.5	48,700	198	2006	681	>0.0	44,500
35	863	412	-14.9	66,600	117	1548	849	-4.1	36,500	199	1711	674	-2.2	44,900
36	712	606	-18.7	48,900	118	1050	577	-11.1	50,800	200	872	424	-14.7	65,000
38	763	694	-17.3	43,800	120	1530	828	-4.3	37,400	201	292	435	<-35.0	63,700
39	304	470	<-35.0	59,800	121	838	423	-15.4	65,200	202	736	253	-18.0	107,800
41	1165	569	-9.2	51,400	122	1572	712	-3.8	42,900	203	786	829	-16.7	37,400
42	684	607	-19.6	48,800	123	23	1433	<-35.0	15,300	204	1224	589	-8.5	50,000
43	1318	589	-7.3	50,000	124	621	1474	-21.9	13,900	205	439	983	-30.9	31,100
44	1924	362	-0.1	74,600	125	1298	862	-7.5	36,000	206	1994	571	>0.0	51,300
46	1203	586	-8.7	50,200	126	872	921	-14.7	33,500	207	1895	687	-0.3	44,200
47	1391	447	-6.3	62,300	127	1000	717	-12.0	42,600	208	240	1418	<-35.0	15,800
48	309	454	<-35.0	61,500	128	1229	311	-8.4	86,100	210	1700	499	-2.3	57,000
49	605	587	-22.5	50,100	129	1422	832	-5.8	37,300	211	902	517	-14.1	55,400
50	621	535	-21.8	53,900	130	1776	499	-1.4	57,000	213	1087	684	-10.4	44,400
51	1113	522	-10.0	55,000	131	1830	757	-0.1	40,700	214	1340	668	-7.0	45,200
52	1820	499	-0.9	57,000	132	660	537	-20.4	53,800	215	1591	495	-3.5	57,300
53	725	177	-18.3	170,800	133	666	1019	-20.2	29,700	216	1585	755	-3.6	40,700
54	2001	500	>0.0	56,900	134	1271	862	-7.9	36,000	217	1159	393	-9.3	69,300
55	722	830	-18.4	37,300	135	1161	1389	-9.3	16,800	218	931	572	-13.5	51,200
56	678	533	-19.8	54,100	136	453	1063	-29.7	28,100	219	713	177	-18.7	170,500
57	1682	302	-2.5	89,000	137	1858	823	-0.6	37,700	220	1479	911	-4.9	33,900
58	1091	580	-10.3	50,600	138	1504	697	-4.6	43,700	221	965	927	-12.8	33,300
59	1171	585	-9.2	50,300	139	1488	707	-4.8	43,200	222	934	716	-13.5	42,700
60	1400	624	-6.2	47,800	140	1689	756	-2.4	40,700	225	1812	1045	-1.0	28,800
61	1853	508	-0.6	56,200	141	311	1417	<-35.0	15,800	226	821	411	-15.8	66,800
62	1888	567	-0.4	51,500	142	1366	915	-6.7	33,800	227	1586	1483	-3.6	13,600
65	735	297	-18.1	90,500	143	1429	346	-5.7	77,900	228	1065	567	-10.8	51,600
66	1263	312	-8.0	85,900	144	615	1017	-22.1	29,800	229	1577	890	-3.7	34,800
67	1252	407	-8.1	67,300	145	2006	566	>0.0	51,600	230	1458	496	-5.2	57,300
68	779	692	-16.8	43,900	146	2006	518	>0.0	55,300	232	1440	849	-5.5	36,500
69	1064	296	-10.8	90,800	147	1070	1108	-10.7	26,500	234	1692	489	-2.4	57,900
71	656	589	-20.6	50,000	148	1347	578	-6.9	50,800	235	618	1004	-22.0	30,300
72	638	545	-21.2	53,100	149	541	1481	-25.7	13,700	236	920	1138	-13.7	25,400
73	1582	583	-3.6	50,400	150	1645	760	-2.8	40,500	237	952	1008	-13.1	30,200
74	1570	556	-3.8	52,300	151	1269	236	-7.9	117,000	238	1611	541	-3.2	53,500
75	1264	621	-8.0	48,000	152	1507	911	-4.5	33,900	239	1489	720	-4.8	42,500
76	1338	564	-7.0	51,800	153	1722	448	-2.1	62,100	240	501	448	-27.7	62,100
77	1833	363	-0.8	74,400	154	932	503	-13.5	56,600	241	1820	569	-0.9	51,400
78	1767	565	-1.5	51,700	155	1031	294	-11.4	91,400	242	1357	658	-6.8	45,800
79	925	738	-13.6	41,600	156	1970	684	>0.0	44,400	243	711	1182	-18.7	23,800
80	534	698	-26.1	43,600	157	1258	183	-8.1	162,400	244	1855	621	-0.6	48,000
81	1811	363	-1.0	74,500	158	1275	417	-7.8	65,900	245	1189	474	-8.9	59,300
82	1412	681	-6.0	44,500	159	1663	820	-2.6	37,800	246	551	459	-25.1	61,000
83	1471	347	-5.0	77,500	160	1034	527	-11.4	54,600	247	1348	604	-6.9	49,100
84	1662	563	-2.7	51,800	161	1953	771	>0.0	40,000	248	460	448	-29.3	62,100
85	1596	479	-3.4	58,900	162	1020	1482	-11.6	13,700	249	1733	451	-1.9	61,800
86	1817	301	-0.9	89,100	164	1566	806	-3.8	38,400	250	1974	788	>0.0	39,200
87	516	1371	-27.0	17,400	166	1905	565	-0.2	51,700	251	808	392	-16.1	69,500
88	1589	698	-3.5	43,600	167	1340	181	-7.0	164,900	252	874	553	-14.6	52,500
89	1706	719	-2.2	42,500	168	1506	583	-4.6	50,400	253	753	848	-17.6	36,500
90	651	329	-20.8	81,700	169	1338	678	-7.0	44,700	254	995	450	-12.1	61,900
91	1415	710	-6.0	43,000	170	1969	541	>0.0	53,500	255	1690	679	-2.4	44,600
92	1773	545	-1.4	53,200	171	800	378	-16.3	71,800	256	994	1006	-12.1	30,200
93	1338	446	-7.0	62,300	172	476	958	-28.7	32,100	257	508	464	-27.4	60,400
94	1708	696	-2.2	43,700	173	919	1314	-13.7	19,300	258	1517	820	-4.4	37,800

Master table of proteins in the rat liver database, showing spot master number, gel position (x and y), isoelectric point relative to CPK standards, and predicted molecular mass (from the standard curve of Fig. 8).

MSN	X	Y	CPKd	SOSMW	MSN	X	Y	CPKd	SOSMW	MSN	X	Y	CPKd	SOSMW	MSN	X
259	1796	961	-1.1	31,800	345	1006	578	-11.9	50,800	426	1296	704	-7.6	43,300	509	800
260	661	1361	-20.4	17,700	346	1095	640	-10.3	46,800	427	810	843	-16.0	36,800	510	1006
261	1725	679	-2.0	44,600	347	625	728	-21.7	42,000	428	1565	303	-3.9	86,700	511	1606
262	486	1127	-28.0	25,800	348	361	963	-35.3	31,100	429	1259	847	-8.0	36,800	512	948
263	1063	172	-10.9	177,400	349	110	1343	-35.0	18,300	430	1253	562	-8.1	51,800	513	481
265	1390	673	-6.3	45,000	350	521	1130	-26.7	25,700	431	734	1426	-18.1	15,300	514	1334
266	510	437	-27.3	63,400	351	912	619	-13.9	48,100	432	483	433	-28.5	63,800	515	868
267	660	1038	-20.4	29,000	352	1574	530	-3.7	54,300	434	518	1041	-26.9	28,800	516	798
268	430	961	-31.0	31,800	353	961	912	-12.9	33,900	435	1020	1170	-11.6	24,300	517	822
269	1044	806	-11.2	48,900	354	706	762	-18.9	40,400	436	1122	196	-9.8	147,800	518	632
270	2019	853	>0.0	36,300	355	1450	830	-5.3	37,300	437	1870	673	-0.5	45,000	519	1332
271	857	422	-15.0	65,200	356	1374	1152	-6.5	24,900	438	435	1102	-31.0	26,700	520	603
272	895	968	-14.2	31,700	357	474	997	-28.7	30,600	439	86	847	-35.0	36,800	521	1190
274	1292	712	-7.6	42,800	358	796	346	-16.3	77,800	440	1740	544	-1.8	53,200	522	479
275	1350	590	-6.9	49,900	359	764	338	-17.3	79,400	441	599	1571	-22.8	10,800	523	768
276	1670	1089	-2.6	27,100	360	1384	1068	-6.4	27,900	442	743	335	-17.8	80,100	524	747
277	688	538	-19.4	53,700	361	1713	789	-2.1	40,100	443	801	668	-16.2	45,200	525	1170
278	961	718	-13.0	42,600	362	1161	859	-9.3	36,100	444	1050	926	-11.1	33,300	526	1502
279	879	570	-14.5	51,300	363	914	1156	-13.8	24,800	445	1245	1298	-8.2	19,800	527	1728
281	1848	1084	-0.7	27,300	364	412	435	-32.0	63,700	446	1576	1516	-3.7	12,800	528	507
282	1505	525	-4.6	54,800	365	741	486	-17.9	58,200	447	1818	1021	-0.9	29,800	529	870
283	1313	1147	-7.3	25,100	366	878	1503	-14.6	13,000	448	1094	440	-10.3	63,100	530	1347
284	1314	829	-7.3	37,400	367	1560	935	-3.9	33,000	449	1945	802	>0.0	38,800	531	1513
285	1332	408	-7.1	67,200	368	963	520	-12.4	55,200	450	1652	894	-2.8	34,800	532	308
286	1277	652	-7.8	46,100	369	434	441	-31.0	63,000	451	1403	500	-6.1	56,800	533	1851
288	1391	824	-6.3	37,600	370	639	610	-21.2	48,700	452	1394	718	-6.3	42,800	534	1483
289	1147	579	-9.5	50,700	371	1587	860	-3.6	36,100	453	905	436	-14.0	63,500	535	809
290	925	511	-13.6	55,900	372	1875	762	-0.5	40,400	454	1038	581	-11.3	50,500	536	625
291	787	1476	-16.6	13,900	373	1351	1059	-6.8	28,300	455	1598	294	-3.4	91,400	537	1164
292	1462	818	-5.1	37,800	374	1506	715	-4.6	42,700	456	1528	863	-4.3	35,800	538	803
293	531	449	-26.3	62,000	375	1823	532	-0.9	54,200	457	1098	1137	-10.2	25,400	539	1259
294	860	698	-14.9	43,600	376	254	417	<-35.0	65,900	458	1814	1072	-0.9	27,800	540	856
295	1162	609	-9.3	48,700	377	1409	583	-6.1	50,400	459	1814	1072	-0.9	27,800	541	803
296	218	814	<-35.0	38,000	378	621	494	-21.8	57,500	460	1388	481	-6.3	58,700	542	1182
297	1377	979	-6.5	31,300	379	1017	595	-11.7	49,400	461	1194	1084	-8.9	27,300	543	128
299	913	1523	-13.9	12,400	380	953	598	-13.1	49,400	462	577	467	-23.9	60,100	544	1355
300	2012	667	>0.0	45,300	381	856	674	-15.0	44,900	463	1140	888	-9.6	34,900	545	895
301	702	178	-19.0	169,200	382	1252	258	-8.1	105,300	464	1293	1133	-7.6	54,800	546	1369
302	494	1280	-28.1	20,400	383	1699	1518	-2.3	12,500	465	618	655	-21.9	46,000	547	982
303	403	1008	-32.6	30,100	384	1042	493	-11.2	57,500	466	2009	299	>0.0	89,900	548	1125
304	1843	1585	-0.7	10,300	385	1490	583	-4.7	50,400	467	1205	215	-8.7	131,300	549	705
305	1049	563	-11.1	49,800	386	1554	603	-4.0	49,100	468	1035	788	-11.4	39,200	550	1477
306	1608	989	-3.3	30,900	387	1193	404	-8.9	67,700	469	160	155	<-35.0	207,800	551	980
307	1219	916	-8.5	33,700	388	1374	902	-6.5	34,300	470	469	1370	-28.9	17,400	552	1028
308	1627	755	-3.0	40,700	389	1456	969	-5.2	31,700	471	599	662	-22.8	45,800	553	888
309	1524	882	-4.4	34,700	390	718	690	-18.5	44,000	472	1009	540	-11.8	53,500	554	788
310	1769	1028	-1.5	29,400	391	1799	732	-1.1	41,900	473	1216	235	-8.6	117,400	555	777
311	1609	1451	-3.3	14,700	392	1482	758	-4.8	40,600	474	816	346	-15.9	77,800	556	980
312	266	1408	<-35.0	16,100	393	1227	1461	-8.4	14,400	475	683	673	-19.3	44,800	557	1519
313	1902	1365	-0.3	17,600	394	1530	577	-4.3	50,800	476	1608	1013	-3.3	30,000	558	1212
314	1316	1395	-7.3	16,800	395	1410	755	-6.0	40,800	477	478	599	-28.6	49,300	559	780
315	1341	523	-7.0	54,900	396	912	256	-13.9	106,400	478	1025	607	-11.5	48,800	560	818
318	1104	1053	-10.1	28,500	397	1465	1063	-5.0	28,100	479	1045	1186	-11.2	23,700	561	1142
320	1480	1459	-4.9	14,400	398	1473	450	-4.9	61,900	480	1609	301	-3.3	89,200	562	532
321	850	603	-15.1	49,100	400	1029	1140	-11.5	25,300	481	775	1289	-17.0	20,100	563	771
322	1454	1494	-5.3	13,300	401	1516	754	-4.4	40,800	482	692	178	-19.3	169,300	564	1068
323	670	626	-20.0	47,700	403	1495	554	-4.7	52,500	483	1100	964	-10.2	31,800	565	822
324	655	101	-20.6	420,500	404	1525	1092	-4.3	27,100	484	1760	776	-1.6	39,700	566	914
325	1521	675	-4.4	44,800	406	723	252	-18.4	108,000	485	882	247	-14.5	110,700	567	1084
326	1587	677	-3.6	44,700	408	650	663	-20.8	45,500	486	494	1436	-28.1	15,200	568	1524
327	1388	408	-6.3	67,000	409	1501	478	-4.6	59,000	487	980	852	-12.5	36,400	569	1392
328	448	1291	-30.0	20,100	410	936	1057	-13.4	28,300	488	1414	546	-6.0	53,100	570	982
330	1608	751	-3.3	40,800	411	350	1120	-35.9	26,000	489	1234	1072	-8.3	27,800	571	1487
331	1566	897	-3.8	43,700	412	1033	538	-11.4	53,700	490	1246	659	-8.2	45,700	572	758
332	531	471	-26.3	59,600	413	737	425	-18.0	64,900	491	824	792	-15.7	39,000	573	887
333	784	1156	-16.7	24,700	415	1578	606	-3.7	48,900	492	1134	1134	-8.2	25,500	574	830
334	1059	407	-10.9	67,300	416	646	496	-21.0	57,300	493	1115	1407	-9.9	16,200	575	1088
335	1593	303	-3.5	88,500	417	1695	482	-2.3	58,600	494	1189	391	-8.9	69,700	576	842
336	1616	598	-3.2	49,400	418	725	770	-18.3	40,000	495	1578	402	-3.7	68,000	577	1317
338	1854	1004	-0.6	30,300	420	1289	1041	-7.7	28,900	496	787	250	-16.6	108,000	578	85
339	1265	888	-8.0	34,900	421	1171	912	-9.1	33,900	497	979	552	-12.5	52,600	579	1014
340	581	585	-23.6	50,300	422	599	162	-22.8	193,700	498	1153	619	-9.4	48,900	580	1827
341	1497	1047	-4.7	28,700	423	929	856	-13.6	36,200	499	1730	1006	-2.0	30,200	581	1009
343	1351	265	-6.8	102,200	424	739	625	-17.9	47,700	500						
344	1813	549	-0.9	52,800	425	1490	965	-4.7	31,800	501						

SN	X	Y	CPKd	SDSMW
511	809	484	-16.0	58,400
512	1099	533	-10.2	54,100
513	1696	1034	-2.3	29,200
514	948	636	-13.2	47,100
515	481	543	-28.5	53,400
516	1334	1044	-7.1	28,800
517	868	1021	-14.8	29,700
518	798	779	-16.3	39,600
519	822	670	-15.7	45,100
520	632	165	-21.5	189,000
521	1332	830	-7.1	37,300
522	603	1104	-22.6	26,600
523	1190	309	-8.9	86,800
524	479	1226	-28.6	22,300
525	768	1066	-17.2	28,000
526	747	1016	-17.7	29,800
527	1170	231	-9.2	119,600
528	1502	542	-4.6	53,400
529	1728	620	-2.0	48,000
530	507	1011	-27.4	30,000
531	870	489	-14.7	57,800
532	1347	1085	-6.9	27,300
533	1513	346	-4.5	77,800
534	308	654	<-35.0	46,000
535	1851	689	-0.7	44,100
536	1463	982	-5.1	31,100
537	909	561	-13.9	52,000
538	625	289	-21.7	83,100
539	1164	198	-9.2	146,200
540	803	655	-16.2	45,900
541	1259	1143	-8.0	25,200
542	856	1526	-15.0	12,200
543	803	1071	-16.2	27,800
544	1162	274	-9.3	98,400
545	128	1321	<-35.0	19,000
546	1355	1122	-6.8	25,900
547	595	866	-23.0	35,800
548	1369	494	-6.6	57,500
549	992	405	-12.2	67,600
550	1125	410	-9.8	66,900
551	705	975	-18.9	31,400
552	1477	1030	-4.9	29,300
553	980	583	-12.5	50,400
554	700	1109	-19.1	26,400
555	1028	621	-11.5	48,000
556	898	794	-14.1	38,900
557	789	1446	-16.6	14,900
558	777	766	-16.9	40,200
559	980	328	-12.5	81,900
560	1519	611	-4.4	48,600
561	1212	661	-8.6	45,600
562	760	594	-17.4	49,700
563	618	956	-21.9	32,100
564	1142	771	-9.6	40,000
565	532	787	-26.2	39,300
566	771	250	-17.1	109,200
567	1088	534	-10.8	54,100
568	822	734	-15.7	41,800
569	914	754	-13.8	40,800
570	1064	794	-10.8	38,900
571	1524	714	-4.4	42,800
572	1392	783	-6.3	39,400
573	982	686	-12.4	44,200
574	1487	672	-4.8	45,000
575	758	731	-17.4	41,900
576	687	1152	-19.5	24,900
577	930	523	-13.5	55,000
578	1888	774	-0.4	39,900
579	642	485	-21.1	58,300
580	1317	519	-7.3	55,300
581	65	1548	<-35.0	11,500
582	1014	814	-11.7	48,400
583	732	176	-18.1	172,300
584	1627	478	-3.0	59,000
585	1009	1426	-11.8	15,500

MSN	X	Y	CPKd	SDSMW
586	619	269	-21.9	100,500
587	1176	461	-9.1	60,700
588	1465	1044	-5.0	28,800
589	741	1188	-17.9	23,600
590	907	402	-14.0	68,000
591	687	658	-19.5	45,800
592	712	1138	-18.7	25,400
593	898	181	-14.1	165,200
594	783	1481	-16.7	14,400
595	736	223	-18.0	125,300
596	629	273	-21.6	98,700
597	1064	286	-10.8	94,000
598	883	503	-14.5	56,700
599	2012	610	>0.0	48,700
600	1255	903	-8.1	34,200
601	1103	391	-10.1	69,600
602	778	265	-16.9	102,000
603	824	518	-15.7	55,400
604	1095	195	-10.3	149,100
605	1759	478	-1.6	56,000
606	994	372	-12.1	72,900
607	751	374	-17.6	72,400
608	1429	518	-5.7	55,300
609	1050	520	-11.1	55,200
610	923	1105	-13.7	26,600
611	1462	622	-5.1	47,900
612	759	225	-17.4	124,000
613	758	1038	-17.4	29,000
614	1438	606	-5.5	48,900
615	1096	1089	-10.2	27,200
616	942	548	-13.3	53,000
617	809	621	-16.0	48,000
618	899	979	-14.1	31,300
619	1135	1321	-9.6	19,100
620	979	615	-12.5	48,300
621	1542	1076	-4.1	27,600
622	1345	814	-6.9	38,000
623	409	950	-32.2	32,400
624	1165	704	-9.2	43,300
625	774	604	-17.0	49,000
626	1263	524	-8.0	54,800
627	952	411	-13.1	66,700
628	1717	575	-2.1	51,000
629	994	292	-12.1	92,000
630	165	1224	<-35.0	22,400
631	803	251	-16.2	108,900
632	719	296	-18.5	90,700
633	1100	294	-10.2	91,400
634	534	1263	-26.1	21,000
635	1153	1038	-9.4	29,000
636	1246	204	-8.2	140,000
637	14	1406	<-35.0	16,200
638	1713	1049	-2.1	28,600
639	1986	1183	>0.0	23,800
640	1378	816	-6.5	38,000
641	1442	1165	-5.5	24,400
642	650	806	-20.8	38,400
643	1111	551	-10.0	52,700
644	1095	861	-10.3	36,000
645	1524	540	-4.4	53,600
646	1777	860	-1.4	36,000
647	391	584	-33.4	50,400
648	977	565	-12.5	51,700
649	658	166	-20.5	187,500
650	732	312	-18.1	86,100
651	1787	567	-1.2	51,500
652	888	268	-14.4	100,900
653	889	775	-14.3	39,800
654	715	221	-18.6	126,300
655	781	227	-16.8	122,400
656	646	165	-21.0	189,100
657	1116	353	-9.9	76,300
658	1382	643	-6.4	46,600
659	547	789	-25.3	39,200
660	984	746	-12.4	41,200

MSN	X	Y	CPKd	SDSMW
674	1661	448	-2.7	62,100
675	1523	562	-4.4	51,800
676	708	642	-18.8	46,700
677	919	615	-13.7	48,300
678	1085	551	-10.5	52,700
679	600	923	-22.7	33,400
680	1237	1004	-8.3	30,300
681	1103	283	-10.1	95,100
682	1406	477	-6.1	59,100
683	1596	249	-3.4	109,800
684	555	699	-24.8	43,500
685	1167	1313	-9.2	19,300
686	1932	790	0.0	39,100
687	1545	619	-4.1	48,100
688	1456	764	-5.2	40,300
689	1011	953	-11.8	32,300
690	1995	270	>0.0	100,200
691	812	888	-16.0	34,900
692	1154	1461	-9.4	14,400
693	1993	819	>0.0	37,800
694	1628	656	-3.0	45,900
695	928	254	-13.6	107,000
696	1854	715	-0.6	42,700
697	1997	345	>0.0	78,000
698	957	563	-13.0	51,800
699	1540	730	-4.2	42,000
700	577	900	-23.8	34,400
701	1610	562	-3.2	51,900
702	1278	571	-7.8	51,200
703	1841	704	-0.7	43,300
704	1018	1386	-11.7	16,900
705	1074	1145	-10.7	25,100
706	293	889	<-35.0	34,800
707	720	412	-18.5	66,600
708	1386	841	-6.4	36,800
709	1328	263	-7.1	103,100
710	698	433	-19.1	63,900
711	701	481	-19.0	58,700
712	1875	699	-0.5	43,600
713	575	702	-23.9	43,400
714	1216	204	-8.6	140,400
715	1069	464	-10.8	60,400
716	1272	506	-7.9	56,400
717	958	822	-13.0	37,700
718	763	395	-17.3	69,100
719	720	916	-18.5	33,700
720	1476	415	-4.9	66,200
721	1846	473	-0.7	59,400
722	510	783	-27.3	39,400
723	1217	1126	-8.6	25,800
724	1858	724	-0.6	42,300
725	665	765	-20.2	40,300
726	1321	312	-7.2	85,900
727	719	427	-18.5	64,600
728	1101	473	-10.2	59,500
729	1359	569	-6.7	51,400
730	696	220	-19.2	127,600
731	687	409	-19.5	67,000
732	1205	256	-8.7	106,200
733	995	563	-12.1	51,900
734	898	596	-14.1	49,500
735	881	181	-14.5	165,900
736	1951	686	>0.0	44,200
737	726	168	-18.3	183,600
738	999	643	-12.0	46,600
739	182	1503	<-35.0	13,000
740	2005	649	>0.0	46,300
741	1448	575	-5.4	51,000
742	792	266	-16.5	101,900
743	469	296	-28.9	90,600
744	664	254	-20.3	107,000
745	1195	184	-8.8	161,000
746	1821	1113	-0.9	26,300
747	909	246	-13.9	111,000
748	790	133	-16.5	264,900

MSN	X	Y	CPKd	SDSMW
761	1399	733	-6.2	41,800
763	1416	1085	-5.9	27,300
764	2020	569	>0.0	51,400
765	651	475	-20.8	59,300
766	1052	1149	-11.1	25,000
767	1968	468	>0.0	59,800
768	1330	685	-7.1	44,300
769	1970	613	>0.0	48,500
770	857	617	-15.0	48,200
771	1337	974	-7.0	31,500
773	1576	502	-3.7	56,700
775	969	824	-12.8	37,600
776	1438	708	-5.5	43,100
777	1539	456	-4.2	61,000
778	850	434	-15.1	63,800
779	700	411	-19.1	66,800
780	1052	1136	-11.1	25,500
784	1413	529	-6.0	54,400
785	1364	885	-6.7	35,000
786	1822	835	-0.9	37,100
787	893	392	-14.3	69,500
790	616	862	-22.0	35,100
791	451	1429	-29.8	15,400
792	777	377	-16.9	72,000
793	1536	1543	-4.2	11,700
794	1461	807	-5.1	38,300
796	388	546	-33.6	53,100
797	1126	212	-9.8	133,700
798	933	437	-13.5	63,400
799	1420	583	-5.9	49,800
800	1759	279	-1.6	96,500
801	624	865	-21.7	35,800
802	898	547	-14.2	53,000
803	1775	1468	-1.4	14,200
804	573	196	-24.0	148,400
805	203	494	<-35.0	57,400
806	980	1039	-12.5	29,000
807	902	308	-14.1	87,200
808	625	827	-21.7	37,500
809	1851	1015	-0.7	29,900
810	440	573	-30.9	51,100
811	1358	249	-6.8	109,700
812	851	393	-15.1	69,400
813	745	1246	-17.8	21,600
814	2028	810	>0.0	38,200
815	1086	645	-10.4	46,500
816	629	313	-21.6	85,700
817	1376	1177	-6.5	24,000
818	1771	790	-1.4	39,100
819	1045	263	-11.2	103,100
820	984	362	-12.4	74,600
821	1712	279	-2.2	96,700
822	1256	205	-8.1	139,200
823	1517	654	-4.4	46,000
824	1442	449	-5.5	62,000
825	1240	513	-8.3	55,800
826	1309	1014	-7.4	29,900
827	2012	708	>0.0	43,100
828	937	1405	-13.4	16,200
830	1342	756	-7.0	40,700
831	562	826	-24.5	37,500
832	1073	1039	-10.7	29,000
833	481	820	-28.5	37,800
834	501	581	-27.8	50,500
837	751	748	-17.6	41,100
838	635	833	-21.3	37,200
839	1494	459	-4.7	60,900
840	1952	301	>0.0	89,300
841	1585	1080	-3.6	27,500
842	571	1312	-24.1	19,400
843	1325	649	-7.2	46,300
844	1727	301	-2.0	89,200
845	630	679	-21.5	44,600
846	2016	905	>0.0	34,200
847	673	1200	-19.9	23,200

MSN	X	Y	CPKd	SDSMW
848	1863	271	-0.6	99,500
849	1166	523	-9.2	54,900
850	1535	1024	-4.2	29,600
851	1035	826	-11.4	37,500
852	834	542	-15.5	53,400
855	499	220	-27.8	127,100
856	1063	194	-10.9	150,500
857	887	890	-14.4	34,800
858	1448	639	-5.4	46,900
859	706	311	-18.9	86,200
860	1070	1066	-10.7	28,000
861	472	347	-28.8	77,600
862	674	480	-19.9	58,800
864	1307	499	-7.4	57,000
865	645	887	-21.0	34,900
866	827	1004	-15.6	30,300
868	685	494	-19.5	57,400
869	1807	402	-1.0	68,000
870	1323	783	-7.2	39,400
871	1228	1031	-8.4	29,300
872	1904	346	-0.3	77,700
873	556	647	-24.8	46,400
874	1540	756	-4.2	40,700
875	1566	777	-3.8	39,700
876	1198	351	-8.8	76,800
877	1076	720	-10.6	42,500
878	1161	1111	-9.3	26,400
879	647	757	-20.9	40,700
880	1756	594	-1.6	49,700
881	1543	278	-4.1	97,100
883	1432	890	-5.7	34,800
884	922	689	-13.7	44,100
885	1103	414	-10.1	66,400
886	1501	607	-4.6	48,900
887	798	1103	-16.3	26,600
888	636	634	-21.3	47,200
889	951	759	-13.1	40,600
890	717	548	-18.6	52,900
891	1123	229	-9.8	121,200
892	891	413	-14.3	66,400
894	1245	234	-8.2	117,800
895	1962	346	>0.0	77,700
896	1322	626	-7.2	47,700
897	420	570	-31.4	51,300
898	662	428	-20.3	64,500
899	845	243	-15.3	113,000
900	624	703	-21.7	43,400
901	931	1094	-13.5	27,000
903	799	229	-16.3	121,000
904	765	520	-17.2	55,200
905	775	889	-17.0	34,800
907	888	824	-14.4	37,600
908	828	1303	-15.6	19,700
910	681	1544	-19.7	11,700
911	1544	301	-4.1	89,100
913	1606	387	-3.3	70,400
914	1237	688	-8.3	44,100
916	1442	749	-5.5	41,100
917	1260	367	-8.0	73,700
919	764	1541	-17.3	11,700
920	1133	1123	-9.7	25,900
921	1123	380	-9.8	71,500
923	829	242	-15.6	113,200
924	1131	318	-9.7	84,300
925	1441	874	-5.5	35,400
926	679	219	-19.7	128,200
927	1487	1191	-4.8	23,500
928	1082	775	-10.5	39,800
929	1231	816	-8.4	38,000
931	1609	670	-3.3	45,100
932	810	900	-16.0	34,400
933	965	520	-12.8	55,100
934	947	462	-13.2	60,600
936	865	843	-14.8	36,800
937	1421	1056	-5.9	28,400

MSN	X	Y	CPKd	SDSMW
939	1197	827	-8.8	37,500
941	1765	885	-1.5	35,000
942	602	472	-22.7	56,600
943	312	498	<-35.0	57,100
944	993	491	-12.1	57,700
945	1300	269	-7.5	100,300
946	630	423	-21.6	65,100
947	187	736	<-35.0	41,600
948	1360	344	-6.5	78,200
949	1766	665	-1.5	45,400
950	1038	183	-11.3	151,000
951	860	152	-14.9	213,000
952	957	701	-13.0	43,400
954	503	547	-27.6	53,000
955	1838	712	>0.0	42,900
957	1010	816	-11.8	37,800
958	768	174	-17.2	174,900
959	596	419	-23.0	65,700
961	557	409	-24.8	83,900
962	887	320	-14.4	80,500
963	564	334	-24.5	24,800
964	969	1155	-12.8	106,600
965	671	255	-20.0	38,700
966	1204	798	-8.7	210,300
967	910	154	-13.9	28,700
968	609	1048	-22.3	138,900
969	1285	206	-7.7	119,300
970	822	232	-15.8	63,400
971	976	437	-12.6	51,600
972	403	567	-32.6	57,400
974	279	495	<-35.0	31,200
975	844	981	-15.3	91,100
976	1124	295	-9.8	45,400
977	994	664	-12.1	46,700
978	1612	642	-3.2	25,300
979	749	1141	-17.7	46,700
980	1064	642	-10.8	33,900
981	1197	911	-8.8	12,800
983	1762	1508	-1.6	84,700
984	1344	317	-6.9	26,600
985	1024	1105	-11.5	24,600
987	739	1159	-17.9	52,400
988	816	555	-15.9	74,900
990	785	361	-16.7	84,500
991	1159	317	-9.3	33,300
992	1090	928	-10.4	43,400
993	1030	701	-11.5	38,200
994	847	811	-15.2	60,700
995	902	461	-14.1	36,600
996	888	847	-14.4	50,700
997	1815	579	-0.9	56,500
998	1205	504	-8.7	93,100
999	617	289	-22.0	92,700
1000	968	290	-12.8	40,000
1001	970	771	-12.7	58,800
1002	1736	478	-1.9	23,700
1003	643	1184	-21.1	58,100
1006	822	487	-15.8	96,400
1007	875	279	-14.6	46,800
1009	291	644	<-35.0	41,200
1010	1386	745	-6.4	53,500
1011	459	541	-29.4	45,600
1012	679	661	-19.7	25,800
1013	1818	1128	-0.9	47,200
1014	1032	634	-11.4	30,700
1015	1629	994	-3.0	25,500
1016	1311	1134	-7.4	65,000
1017	1722	424	-2.0	41,300
1018	1015	743	-11.7	22,900
1020	1574	1219	-3.7	58,400
1021	781	484	-16.8	581,300
1022	1129	83	-9.7	84,800
1023	812	317	-15.9	62,400
1024	785	446	-16.7	41,500
1025	1290	739	-7.7	

MSN	X	Y	CPKd	SDSMW	MSN	X	Y	CPKd	SDSMW	MSN	X	Y	CPKd	SDSMW
1026	405	552	-32.3	52,600	1153	921	1158	-13.7	24,700	1246	547	577	-25.3	50,800
1027	1298	848	-7.5	36,500	1154	1584	864	-3.5	35,900	1247	530	576	-26.3	50,900
1028	856	547	-15.0	53,000	1161	637	400	-21.3	68,400	1249	516	572	-27.0	51,200
1030	1284	226	-7.7	123,200	1162	623	397	-21.8	68,800	1250	973	536	-12.7	53,900
1031	886	822	-12.3	37,700	1163	665	397	-20.2	68,700	1251	607	532	-22.4	54,200
1032	1547	403	-4.1	67,900	1168	564	528	-24.4	54,500	1252	665	529	-20.2	54,400
1033	1381	551	-6.4	52,700	1170	552	529	-25.0	54,500	1253	899	766	-14.1	40,200
1034	1525	486	-4.3	57,200	1171	538	524	-25.9	54,800	1254	1311	746	-7.4	41,200
1035	1128	645	-9.7	46,500	1172	545	514	-25.5	55,700	1255	1300	761	-7.5	40,400
1036	1226	274	-8.5	98,300	1174	1099	522	-10.2	55,000	1257	1938	712	0.0	42,900
1039	1781	262	-1.6	103,600	1176	1304	586	-7.5	50,200	1258	1806	718	-1.0	42,600
1040	541	839	-25.7	36,900	1177	1366	539	-6.6	53,700	1259	1727	715	-2.0	42,700
1041	818	910	-15.8	34,000	1178	1608	702	-3.3	43,400	1260	1629	713	-3.0	42,800
1044	1036	485	-11.3	58,300	1179	1485	224	-4.8	124,900	1261	1555	717	-4.0	42,600
1045	1439	407	-5.5	67,300	1180	1459	224	-5.2	124,900	1262	1468	717	-5.0	42,600
1047	1540	250	-4.2	109,200	1181	1431	223	-5.7	125,100	1263	1413	722	-6.0	42,400
1048	1576	635	-3.7	47,100	1182	1407	223	-6.1	125,200	1264	1340	717	-7.0	42,600
1049	1089	411	-10.4	66,700	1183	1383	224	-6.4	124,700	1265	1263	717	-8.0	42,600
1050	949	1040	-13.2	28,900	1184	1454	182	-5.3	164,400	1266	1182	720	-9.0	42,500
1051	426	818	-31.1	37,800	1185	1422	183	-5.8	162,600	1267	1110	717	-10.0	42,600
1052	1583	1385	-3.6	16,900	1186	1394	182	-6.3	164,300	1268	1055	717	-11.0	42,600
1053	779	1092	-16.8	27,000	1189	1171	214	-9.2	131,800	1269	999	717	-12.0	42,600
1054	1613	620	-3.2	48,000	1190	1457	286	-5.2	94,200	1270	959	715	-13.0	42,700
1055	1380	377	-6.5	72,000	1191	686	1114	-19.5	26,200	1271	905	712	-14.0	42,900
1056	284	663	<-35.0	45,500	1192	265	893	<-35.0	34,700	1272	857	714	-15.0	42,800
1058	1261	746	-8.0	41,200	1193	403	1292	-32.6	20,000	1273	810	705	-16.0	43,300
1060	393	605	-33.3	49,000	1194	344	1275	<-35.0	20,600	1274	774	711	-17.0	42,900
1061	1817	645	-0.9	46,600	1195	505	1311	-27.6	19,400	1277	737	708	-18.0	43,100
1062	1245	746	-8.2	41,200	1196	572	1293	-24.1	20,000	1278	702	711	-19.0	42,900
1064	1258	792	-8.1	39,000	1197	639	1502	-21.2	13,000	1279	671	710	-20.0	43,000
1065	705	934	-18.9	33,000	1198	637	1402	-21.3	16,300	1280	645	710	-21.0	43,000
1066	1181	734	-9.0	41,800	1199	614	1407	-22.1	16,200	1281	617	707	-22.0	43,100
1067	529	658	-26.3	45,800	1200	637	1431	-21.3	15,400	1282	595	704	-23.0	43,300
1068	508	686	-27.4	43,700	1201	1095	1394	-10.3	16,600	1283	573	700	-24.0	43,500
1069	1898	604	-0.3	49,100	1202	1719	1545	-2.1	11,600	1284	552	695	-25.0	43,700
1071	873	609	-14.7	48,700	1203	791	668	-16.5	45,200	1285	536	694	-26.0	43,800
1073	1768	1128	-1.5	25,800	1204	964	1021	-12.9	29,700	1286	515	687	-27.0	44,200
1075	836	773	-15.4	39,900	1205	313	195	<-35.0	148,700	1287	496	683	-28.0	44,400
1076	1863	861	-0.6	36,000	1208	306	194	<-35.0	149,800	1288	467	669	-29.0	45,200
1078	826	566	-15.7	51,600	1209	320	197	<-35.0	147,400	1289	447	667	-30.9	45,300
1081	971	483	-12.7	58,500	1210	326	197	<-35.0	146,600	1290	427	655	-31.0	45,900
1083	1697	202	-2.3	142,300	1211	394	294	-33.2	91,400	1291	412	655	-32.0	45,900
1085	1157	794	-9.4	38,900	1212	402	294	-32.7	91,200	1292	397	652	-33.0	46,100
1090	620	910	-21.9	34,000	1214	386	294	-33.7	91,400	1293	381	654	-34.0	46,000
1092	1867	597	-0.5	49,500	1215	641	329	-21.2	81,600	1294	365	653	-35.0	46,100
1093	2019	894	>0.0	34,600	1216	660	329	-20.4	81,600	1295	348	653	<-35.0	46,100
1094	1546	538	-4.1	53,700	1217	914	266	-13.8	101,800					
1095	1545	477	-4.1	59,100	1218	873	245	-14.7	112,000					
1098	61	935	<-35.0	33,000	1219	970	372	-12.7	72,900					
1099	1954	237	>0.0	116,000	1220	1021	298	-11.6	90,100					
1101	588	1048	-23.3	28,600	1221	1392	205	-6.3	139,500					
1102	1050	667	-11.1	45,200	1222	1354	203	-6.8	141,800					
1103	457	797	-29.5	38,800	1223	1362	205	-6.7	139,500					
1105	1884	532	-0.4	54,200	1224	673	540	-19.9	53,600					
1106	1714	649	-2.1	46,300	1225	614	542	-22.1	53,400					
1107	1717	546	-2.1	53,100	1226	603	539	-22.6	53,600					
1108	1976	722	>0.0	42,400	1227	696	623	-19.2	47,800					
1111	547	1066	-25.3	28,000	1228	707	628	-18.9	47,500					
1112	1348	621	-6.9	48,000	1229	475	447	-28.7	62,300					
1115	1385	762	-6.4	40,400	1230	466	1282	-29.0	20,400					
1116	1078	816	-10.6	38,000	1231	759	1461	-17.4	14,400					
1117	975	787	-12.6	39,300	1232	1324	1170	-7.2	24,200					
1118	1202	933	-8.7	33,100	1233	1583	1005	-3.6	30,300					
1119	1022	1076	-11.6	27,600	1234	1865	809	-0.6	38,200					
1120	1905	816	-0.3	48,300	1235	1812	817	-1.0	37,900					
1121	1512	1301	-4.5	19,700	1236	1411	703	-6.0	43,400					
1122	1114	677	-9.9	44,700	1237	1392	682	-6.3	44,500					
1123	1464	452	-5.1	61,700	1238	794	410	-16.4	66,900					
1125	1048	857	-11.1	36,200	1239	769	407	-17.1	67,300					
1126	1122	802	-9.8	38,600	1240	740	406	-17.9	67,500					
1128	1722	892	-2.1	34,700	1241	743	511	-17.8	55,900					
1133	1098	825	-10.2	37,500	1242	713	510	-18.7	56,000					
1139	1830	569	-0.8	51,400	1243	682	509	-18.6	56,100					
1147	764	1182	-17.3	23,800	1244	663	504	-20.3	56,500					
1148	1968	724	>0.0	42,300	1245	565	582	-24.4	50,500					

Table 2. Table of some identified spots

Protein name	Spot number	Protein	Reference
IDS:3_ALPHA_HDDH	137, 159	Pure protein and antibody provided by Dr. T.M. Penning, Department of Pharmacology, School of Medicine, University of Pennsylvania.	
IDS:ACTIN_BETA	38	Homologous position with respect to other mammalian systems	
IDS:ACTIN_GAMMA	68	Homologous position with respect to other mammalian systems	
IDS:ALBUMIN	21, 28, 33	Predominance in rat plasma	
IDS:APO_A-I	236, 463	Presence in rat plasma, regulation by some lipid-lowering drugs	
IDS:CALMODULIN	123, 649	Homologous position with respect to other mammalian systems	
IDS:CATALASE	54, 61, 106	Presence in purified peroxisomes, similarity in position to mouse catalase	
IDS:CPKSPOTS	1257 - 1295		
IDS:CPS	114, 157, 167, 174, 1184, 1185, 1186, 1222		
IDS:CYTOCHROME_B5	87, 477	Pure protein provided by Dr. Margaret Marshall, Department of Pharmacology, Medical School, University of Wisconsin - Madison.	
IDS:FABP-L	227	Pure protein provided by Dr. Andrew Parkinson, Department of Pharmacology, Toxicology and Therapeutics, University of Kansas Medical Center	
IDS:HMG-COA_SYNTHASE		Pure protein provided by Dr. Nathan Bass, Department of Medicine, University of California School of Medicine, San Francisco	
IDS:LAMIN_B	133, 144, 235, 413	Antibody provided by Dr. Michael Greenspan, Merck Sharp & Dohme Research Laboratories, Rahway, NJ	
IDS:MITCON:1	415, 734	Homologous position with respect to other mammalian systems	
IDS:MITCON:2	17, 49, 71, 340, 1245, 1246, 1247, 1249	Homologous position with respect to other mammalian systems, presence in mitochondria	
IDS:MITCON:3	15, 25, 110, 1241, 1242, 1243, 1244	Homologous position with respect to other mammalian systems, presence in mitochondria	
IDS:NADPH_P450_RED	18, 35, 226, 600, 1238, 1239, 1240	Homologous position with respect to other mammalian systems, presence in mitochondria	
IDS:PD1	175, 251, 812	Pure protein provided by Dr. Andrew Parkinson, Department of Pharmacology, Toxicology and Therapeutics, University of Kansas Medical Center	
IDS:PLASMA_PROTEINS	168, 1170, 1171, 1172	Sequence information obtained by R.M. Van Frank, Lilly Research Laboratories, Indianapolis	
IDS:PRO-ALBUMIN	21, 28, 33, 44, 72, 102, 115, 197, 236, 246, 248, 257, 293, 332, 347, 364, 369, 419, 432, 463, 488, 518, 562, 605, 623, 666, 667, 725, 738, 790, 865, 903, 926	Plasma coelectrophoresis studies	
IDS:PYRCARBOX			
IDS:SOD	179, 1180, 1181, 1182, 1183	Relative position to mature albumin, presence in microsomes	
IDS:TUBULIN_ALPHA	56, 132, 1224, 1252	Pavlica, R.J., et al. BBA (1990) 1022 115-125. Sequence information obtained by R.M. Van Frank, Lilly Research Laboratories, Indianapolis	
IDS:TUBULIN_BETA	50, 1225, 1226, 1251	Homologous position with respect to other mammalian systems	

Hb-beta

Computed
hemoglobinProtein
Rabbit

3. Computed pI's of two sets of carbamylated protein standards: Rabbit muscle CPK and human hemoglobin (Hb)

Protein Name	PIR Name	#ASP 3.9	#GLU 4.1	#HIS 6.0	#LYS 10.8	#ARG 12.5	NH2- 7.0	Calc pI	Real CPK
Rabbit muscle CPK	KIRBCM	28	27	17	34	18	1	6.84	0.0
		28	27	17	33	18	1	6.67	-1
		28	27	17	32	18	1	6.54	-2
		28	27	17	31	18	1	6.42	-3
		28	27	17	30	18	1	6.31	-4
		28	27	17	29	18	1	6.21	-5
		28	27	17	28	18	1	6.12	-6
		28	27	17	27	18	1	6.03	-7
		28	27	17	26	18	1	5.94	-8
		28	27	17	25	18	1	5.85	-9
		28	27	17	24	18	1	5.76	-10
		28	27	17	23	18	1	5.67	-11
		28	27	17	22	18	1	5.58	-12
		28	27	17	21	18	1	5.48	-13
		28	27	17	20	18	1	5.39	-14
		28	27	17	19	18	1	5.29	-15
		28	27	17	18	18	1	5.20	-16
		28	27	17	17	18	1	5.12	-17
		28	27	17	16	18	1	5.04	-18
		28	27	17	15	18	1	4.96	-19
		28	27	17	14	18	1	4.89	-20
		28	27	17	13	18	1	4.83	-21
		28	27	17	12	18	1	4.77	-22
		28	27	17	11	18	1	4.71	-23
		28	27	17	10	18	1	4.66	-24
		28	27	17	9	18	1	4.61	-25
		28	27	17	8	18	1	4.56	-26
		28	27	17	7	18	1	4.52	-27
		28	27	17	6	18	1	4.48	-28
		28	27	17	5	18	1	4.44	-29
		28	27	17	4	18	1	4.40	-30
		28	27	17	3	18	1	4.36	-31
		28	27	17	2	18	1	4.32	-32
		28	27	17	1	18	1	4.29	-33
		28	27	17	0	18	1	4.25	-34
		28	27	17	0	18	0	4.22	-35
Hb-beta, human	HBHU	7	8	9	11	3	1	7.18	
		7	8	9	10	3	1	6.79	
		7	8	9	9	3	1	6.53	-1.8
		7	8	9	8	3	1	6.32	-3.2
		7	8	9	7	3	1	6.13	-5.3
		7	8	9	6	3	1	5.96	-7.2
		7	8	9	5	3	1	5.78	-10.0
		7	8	9	4	3	1	5.59	-12.3
		7	8	9	3	3	1	5.37	-15.5
		7	8	9	2	3	1	5.14	-18.0
		7	8	9	1	3	1	4.91	-21.0
		7	8	9	0	3	1	4.71	-25.5
		7	8	9	0	3	0	4.54	-27.2

Table 4. Computed pI's of some known proteins related to measured CPK pI's

Protein Name	PIR Name	#ASP 3.9	#GLU 4.1	#HIS 6.0	#LYS 10.8	#ARG 12.5	Calc pI	Real CPK
0 Creatine phospho kinase (CPK), rabbit muscle	KIRBCM	28	27	17	34	18	6.84	0.0
1 Fatty acid-binding protein, rat hepatic	FZRTL	5	13	2	16	2	7.83	-3.0
2 b2-microglobulin, human	MGHUB2	7	8	4	8	5	6.09	-5.0
3 Carbamoyl-phosphate synthase, rat	SYRTCA	72	96	28	95	56	5.97	-5.5
4 Proalbumin (serum albumin precursor), rat	ABRTS	32	57	15	53	27	5.98	-6.2
Serum albumin, rat	ABRTS	32	57	15	53	24	5.71	-9.0
Superoxid dismutase (Cu-Zn, SOD), rat	A26810	8	11	10	9	4	5.91	-9.2
Phospholipase C, phosphoinositide-specific (?), rat	A28807	34	42	9	49	21	5.92	-9.2
Albumin, human	ABHUS	36	61	16	60	24	5.70	-11.9
Apo A-I lipoprotein, rat	A24700	18	24	6	23	12	5.32	-13.7
proApo A-I lipoprotein, human	LPHUA1	16	30	6	21	17	5.35	-14.3
NADPH cytochrome P-450 reductase, rat	RDRT04	41	60	21	38	36	5.07	-15.6
Retinol binding protein, human	VAHU	18	10	2	10	14	5.04	-16.9
Actin beta, rat	ATRTC	23	26	9	19	18	5.06	-17.2
Actin gamma, rat	ATRTC	20	29	9	19	18	5.07	-16.6
Apo A-I lipoprotein, human	LPHUA1	16	30	5	21	16	5.10	-17.5
Apo A-IV lipoprotein, human	LPHUA4	20	49	8	28	24	4.88	-19.7
Tubulin alpha, rat	UBRTA	27	37	13	19	21	4.66	-19.8
F1ATPase beta, bovine	PWBOB	25	36	9	22	22	4.80	-21.0
Tubulin beta, pig	UBPGB	26	36	10	15	22	4.49	-22.5
Protein disulphide isomerase (PDI), rat hepatic	ISRTSS	43	51	11	51	9	4.07	-25.0
Cytochrome b5, rat	CBRT5	10	15	6	10	4	4.59	-26.0
Apo C-II lipoprotein, human	LPHUC2	4	7	0	6	1	4.44	-30.5
Amino acid pI assumed in calculation:		3.9	4.1	6.0	10.8	12.5		

Wirth

Luo

Fujimoto

C. Bisgaard

D. Olson

atory of Exp.

ogenesis.

Cancer In

Institutes

oda,

oda,

oda,

oda,

oda,

oda,

oda,

oda,

oda,

oda,

oda,

oda,

oda,

oda,

oda,

oda,

oda,

oda,

oda,

oda,

oda,

oda,

oda,

oda,

oda,

oda,

oda,

oda,

oda,

oda,

oda,

oda,

oda,

oda,

oda,

oda,

oda,

oda,

oda,

oda,

oda,

oda,

oda,

oda,

oda,

oda,

oda,

oda,

oda,

oda,

oda,

oda,

oda,

oda,

oda,

oda,

oda,

oda,

oda,

oda,

oda,

oda,

oda,

oda,

oda,

N. Leigh Anderson¹
Ricardo Esquer-Blasco¹
Jean-Paul Hofmann¹
Lydie Meheus²
Jos Raymackers³
Sandra Steiner⁴
Frank Witzmann⁴
Norman G. Anderson¹

¹Large Scale Biology Corporation,
Rockville, MD

²Innogenetics NV, Ghent

³Sandoz Pharma Ltd, Drug Safety
Assessment, Toxicology, Basel

⁴Molecular Anatomy Laboratory,
Indiana University Purdue

University Columbus, Columbus, IN

An updated two-dimensional gel database of rat liver proteins useful in gene regulation and drug effect studies

We have improved upon the reference two-dimensional (2-D) electrophoretic map of rat liver proteins originally published in 1991 (N. L. Anderson *et al.*, *Electrophoresis* 1991, 12, 907-930). A total of 53 proteins (102 spots) are now identified, many by microsequencing. In most cases, spots cut from wet, Coomassie Blue stained 2-D gels were submitted to internal tryptic digestion [2], and individual peptides, separated by high-performance liquid chromatography (HPLC), were sequenced using a Perkin-Elmer 477A sequenator. Additional spots were identified using specific antibodies.

Figure 1 shows the current annotated 2-D map of F344 rat liver, analyzed using the Iso-DALT system (20 × 25 cm gels) and BDH 4-8 carrier ampholytes. Both the map itself and the master spot number system remain the same as shown in the original publication. Table 1 lists the important features of each identification shown, including the gel position, *pI*, and *M_r*, for the most abundant or most basic form of each protein. Using this extended base of identified spots, a series of four improved calibration functions has been derived for the *pI* and SDS-*M_r* axes (the first two of which are shown in Fig. 2A and B). Both forward and reverse functions are derived, so that one can compute the physical properties of a spot with a given gel location, or inversely compute the gel position expected for a protein having given physical properties:

$$Y_{\text{RATLIVER}} = f_{M-\text{RATLIVER}}(M_{\text{SEQUENCE-DERIVED}}) \quad (1)$$

$$X_{\text{RATLIVER}} = f_{pI-\text{RATLIVER}}(pI_{\text{SEQUENCE-DERIVED}}) \quad (2)$$

$$M_{\text{GEL-DERIVED}} = f_{\text{RATLIVER } Y-M}(Y_{\text{RATLIVER}}) \quad (3)$$

$$pI_{\text{GEL-DERIVED}} = f_{\text{RATLIVER } X-pI}(X_{\text{RATLIVER}}) \quad (4)$$

A spreadsheet program (in Microsoft Excel) was developed to facilitate flexible computation of *pI*'s from amino acid sequence data, and the results were entered into a relational database (Microsoft Access). A table of spot positions and sequence-derived *pI*'s and *M_r*'s was fitted with a large series of analytic equations using Tablecurve (Jandel Scientific), and the four conversion Eqs. (1)-(4), relating computed *pI* and gel *X* coordinate, or computed molecular weight and gel *Y* coordinate, were selected, based on criteria of simplicity, goodness of fit and favorable asymptotic behavior. Table 2 lists the equations and coefficients. Application of Eqs. (3) and (4) to a spot's *X* and *Y* coordinates, given in [1], produce improved *M_r* estimates, and allow computation of *pI*

directly in *pH* units, instead of in terms of positions relative to creatine phosphokinase (CPK) charge standards. The inverse Eqs. (1) and (2) were used to compute the gel positions of a series of *pI* and *M_r* tick marks. These tick marks were plotted with SigmaPlot (Jandel), together with fiducial marks locating several prominent spots, and the resulting graphic was aligned over the synthetic gel image (computed by Kepler from the master gel pattern) using Freelance (Lotus Development). Maps were printed as Postscript output from Freelance, either in black and white (as shown here) or in color, where label color indicates subcellular location (available from the first author upon request). We have also used the rat liver 2-D pattern as presented here to calibrate the patterns of other samples. Using mixtures of rat liver and mouse liver samples, for example, we made composite 2-D patterns that allow use of the rat pattern to standardize both axes of the mouse pattern. This was accomplished by deriving transformations relating the rat and mouse *X*, and separately the rat and mouse *Y*, axes (Table 2, lower half; Fig. 2C and D) based on a series of spots that coelectrophorese in these closely related species. These functions were then applied to derive equations relating the mouse liver *X* and *Y* to *pI* and SDS-*M_r* (Eqs. 5 and 6 below). The resulting standardized 2-D pattern for B6C3F1 mouse liver is shown in Fig. 3.

$$M_{\text{MOUSELIVER}} = f_{\text{RATLIVER } Y-M}(f_{\text{MOUSELIVER } Y-\text{RATLIVER } Y}(Y_{\text{MOUSELIVER}})) \quad (5)$$

$$pI_{\text{MOUSELIVER}} = f_{\text{RATLIVER } X-pI}(f_{\text{MOUSELIVER } X-\text{RATLIVER } X}(X_{\text{MOUSELIVER}})) \quad (6)$$

A slightly more complex approach can be used to standardize samples that have few or no spots co-electrophoresing with rat liver proteins. In this case, a 2-D gel is prepared with a mixture of the two samples, and four functions (forward and backward, each for *X* and *Y*) are derived relating each sample's own master pattern to the composite. The required functions are then applied in a nested fashion to yield the desired result (using rat plasma as an example):

$$M_{\text{RATPLASMA}} = f_{\text{RATLIVER } Y-M}(f_{\text{RATPLASMA-LIVER } Y-\text{RATLIVER } Y}(f_{\text{RATPLASMA } Y-\text{RATPLASMA-LIVER } Y}(Y_{\text{RATPLASMA}})))) \quad (7)$$

Correspondence: Dr. Leigh Anderson, Large Scale Biology Corporation, 9620 Medical Center Drive, Rockville, MD 20850-3338 USA (Tel: +301-424-5989; Fax: +301-762-4892; email: leigh@lsbc.com)

Keywords: Two-dimensional polyacrylamide gel electrophoresis / Liver / Map / Identification / Calibration

Table 1. continued

MSN ^{a)}	Protein IDb)	Protein name	Identification comments	Gel X ^{c)}	Experimental pI ^{d)}	Gel Y ^{c)}	Experimental M _r ^{d)}
1184, 1186, 114, 174, 118, 5, 167, 157, 54, 61, 136	CPSM_RAT CATA_RAT COX2_RAT	Carbamyl phosphate synthase Catalase COX-II	2-D of pure protein; confirmed by N-terminal sequence and AAA Internal sequence Ab (J. W. Taanman), confirmed by internal sequence	1453.56 2000.81 452.57	6.05 6.73 4.61	181.64 499.64 1062.67	160 640 58 968 25 504
87	CYB5_RAT	Cytochrome B5	2-D of pure protein; Ab; confirmed by AAA	515.68	4.73	1370.55	18 493
41	CK-RAT ^{a)}	Cytokeratin	Location in cytoskeletal fraction	1165.12	5.75	569.09	51 448
29	CK-RAT ^{a)}	Cytokeratin	Location in cytoskeletal fraction	743.11	5.15	605.23	48 187
5, 11	ENPL-RAT ^{a)}	Endoplasmin	Ab (F. Witzmann)	567.73	4.83	263.37	112 194
60	ENOA_RAT	Enolase A	Internal sequence and AAA	1399.78	6.00	623.54	46 674
27	ER60_RAT	ER-60	N-Terminal sequence (R. M. Van Frank)	1184.20	5.77	523.51	56 169
17	ATPB_RAT	F1 ATPase β	N-Terminal sequence and AAA	629.06	4.95	588.83	49 620
196	ATP7_RAT	F1 ATPase δ	Internal sequence	1227.24	5.82	1184.65	22 310
79	F16P_RAT	Fructose-1,6-bis-phosphatase	Uncertain; by comparison with ID in Garrison and Wager (JBC 257:13135-13143)	924.54	5.44	737.77	38 858
62, 78, 125	DHE3_RAT HAST-RAT ^{a)}	Glutamate dehydrogenase HAST-I: N-hydroxyaryl-amine sulfotransferase	N-Terminal sequence and internal sequence Internal sequence	1887.39 1297.94	6.55 5.89	566.92 861.55	51 655 32 638
307	HO1_RAT	Heme oxygenase 1	Uncertain; available data from internal sequence	1219.39	5.81	915.71	30 423
413, 1250, 933	HMCS_RAT	HMG CoA synthase, cytosolic	Ab (J. Gernershausen)	1033.48	5.59	538.13	54 571
133, 144, 235	HMCS_RAT	HMG CoA synthase, mitochondrial (frag)	Ab (J. Gernershausen), N-terminal sequence (Steiner/Lottspeich)	666.40	5.02	1019.42	26 811
8, 23, 1307	HS7C_RAT	HSC-70	Positional homology (with human, etc.) through coelectrophoresis	811.87	5.27	425.76	69 521
15, 25, 110	P60_RAT	HSP-60	Ab (F. Witzman); confirmed by N-terminal sequence and AAA	845.09	5.32	520.03	56 561
971	HS70-RAT ^{a)}	HSP-70	Ab (F. Witzman)	976.11	5.51	437.14	67 674
1216, 1215, 90	HS90-RAT ^{a)}	HSP-90	Ab (F. Witzman)	659.86	5.00	329	90 107
256	ING1-HUMAN	Interferon- γ induced protein	Internal sequence	993.85	5.54	1006.04	27 237
415, 734	LAMB-RAT ^{a)}	Lamin B	Positional homology with human through coelectrophoresis, nuclear location	737.10	5.14	425.19	69 615
80	LAMR-RAT ^{a)}	"Laminin receptor"	Internal sequence	534.02	4.77	697.62	41 327
227	FABL_RAT	L-FABP (liver fatty acid binding protein)	Ab (N. M. Bass)	1586.09	6.18	1483.43	16 622
134	MDHC_MOUSE E	Malate dehydrogenase E	Internal sequence	1270.85	5.86	861.96	32 620
18, 35, 226	GR75-RAT ^{a)}	Mitcon-3; grp75	Positional homology with human through coelectrophoresis	905.67	5.41	413.67	71 589
175, 251	NCPR_RAT	NADPH P450 reductase	2-D of pure protein	824.69	5.29	393.21	75 366
1168, 1170, 1171	PDI_RAT	PDI: Protein disulfide isomerase	N-Terminal sequence (R. M. van Frank), Ab	564.30	4.83	528.47	55 618
47, 93	ALBU_RAT	Pro-Albumin	Microsomal lumen location, pI, M _r relative to albumin	1391.03	5.99	446.68	66 195
236	APA1_RAT	Pro-APO A-I lipoprotein	Coelectrophoresis with plasma protein	920.41	5.43	1137.51	23 467
320	IPK1_BOVIN	Protein kinase C inhibitor 1	Internal sequence; homology with bovine protein	1480.01	6.08	1458.81	17 007
152	PNPH_MOUSE	Purine nucleoside phosphorylase	Internal sequence	1507.19	6.10	911.16	30 599
1179, 1180, 1181, 1182, 1183	PYVC-RAT ^{a)}	Pyruvate carboxylase	Tentative; 2-D of pure protein (J. G. Henslee, JBC, 1979); reported in <i>Biochim. Biophys. Acta</i> 1022, 115-125	1485.10	6.08	223.52	131 589
55, 103	SM30_RAT	SMP-30: Senescence marker protein-30	Internal sequence	721.71	5.11	830.10	34 051
135	SODC_RAT	Superoxide dismutase	AAA; confirmed by internal sequence (R. M. Van Frank)	1161.24	5.74	1388.68	18 173
172	TPM-RAT ^{a)}	Tim: tropomyosin	Location in cytoskeleton, 2-D position relative to human, Ab	476.24	4.66	957.86	28 865
277, 56	TBA1_RAT	Tubulin α	Positional homology with human through coelectrophoresis, cytoskeletal location	688.22	5.06	537.67	54 620
50, 1225	TBB1_RAT	Tubulin β	Positional homology with human through coelectrophoresis, cytoskeletal location	621.29	4.93	535.48	54 855
1224	VIME_RAT	Vimentin	Positional homology with human through coelectrophoresis, cytoskeletal location	673.00	5.03	539.50	54 426

B6C3F1 MOUSE LIVER 2-D PROTEIN PATTERN

v.1.1 (B6C3F1MST2) 28-Apr-1995 © by Large Scale Biology Corporation,
9430 Medical Center Drive, Rockville, MD 20850 USA 301/434-3999
MW and computed pI scales shown at top and left of protein pattern

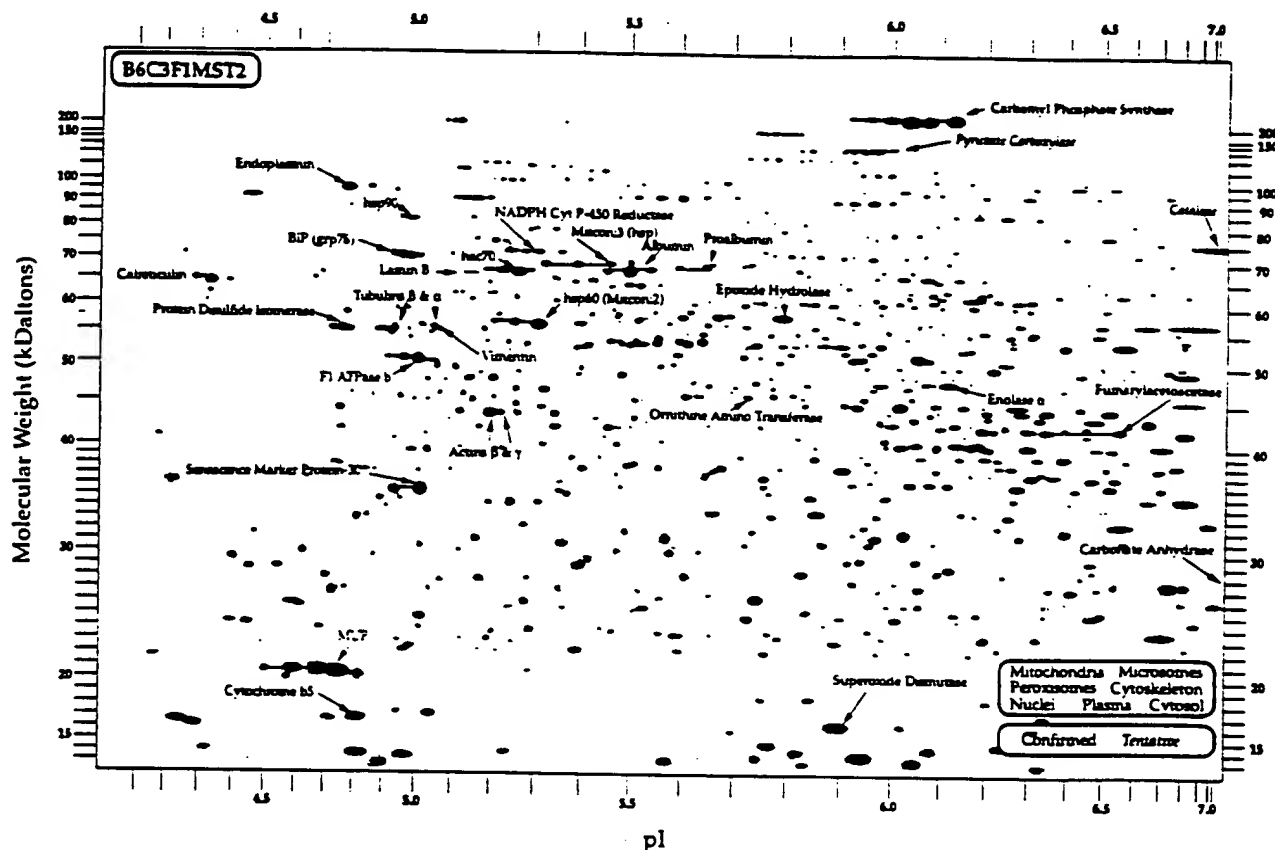


Figure 3. Master 2-D gel pattern for B6C3F1 mouse liver, standardized using the F344 rat liver pattern identifications, according to the method described in the text. Twenty-nine proteins are identified.

$$pI_{\text{RATPLASMA}} = \frac{\int_{\text{RATLIVER}} X \cdot pI \cdot \int_{\text{RATPLASMA-LIVER}} X \cdot \int_{\text{RATPLASMA}} X}{\int_{\text{RATPLASMA-LIVER}} X \cdot \int_{\text{RATPLASMA-LIVER}} X \cdot \int_{\text{RATPLASMA}} X} \quad (8)$$

This unified approach, in which one well-populated 2-D pattern is used to standardize a family of other patterns, has the additional advantage that the resulting pI and M_r scales are directly compatible. Hence one can compare the relative pI 's of mouse and rat versions of a sequenced protein in a consistent pI measurement system, and select likely inter-species analogs based on positional relationships on common scales. Adoption of immobilized pH gradient (IPG) technology [4-7] will result in substantial improvements in pI positional reproducibility for standard 2-D maps such as those presented here; however, we believe that our approach will continue to be useful in establishing the empirical pH gradient actually achieved by such gels under given experimental conditions (temperature, urea concentration, etc.), in relating patterns run on different IPG ranges and using different lots of IPG gels (between which some variation will persist). Development of rodent organ maps is a continuing effort in our laboratories [8-10], and results in regular additions of identified proteins. Those who wish to receive current rodent liver maps, with color annotations, should send a stamped self-addressed envelope to the first author.

We would like to thank the individuals who provided antibodies mentioned in Table 1, and R. M. van Frank for unpublished sequenced data.

Received May 31, 1995

References

- [1] Anderson, N. L., Esquer-Blasco, R., Hofmann, J.-P., Anderson, N. G., *Electrophoresis* 1991, 12, 907-930.
- [2] Rosenfeld, J., Capdevielle, J., Guillemot, J. C., Ferrara, P., *Anal. Biochem.* 1992, 203, 173-179.
- [3] Witzmann, F., Clack, J., Fultz, C., Jarnot, B., *Electrophoresis* 1995, 16, 451-459.
- [4] Rosengren, A. E., Bjellqvist, B., Gasparic, V., *US Patent* 4130470, December 1978.
- [5] Gianazza, E., Artoni, G., Righetti, P. G., *Electrophoresis* 1983, 4, 321-326.
- [6] Görg, A., Postel, W., Günther, S., Weser, J., *Electrophoresis* 1985, 6, 599-604.
- [7] Gianazza, E., Astrua-Testori, S., Giacon, P., Righetti, P. G., *Electrophoresis* 1985, 6, 332-339.
- [8] Myers, T. G., Dietz, E. C., Anderson, N. L., Khairallah, E. A., Cohen, S. D., Nelson, S. D., *Chem. Res. Toxicol.* 1995, 8, 403-413.
- [9] Cunningham, M. L., Pippin, L. L., Anderson, N. L., Wenk, M. L., *Toxicol. Appl. Pharmacol.* 1995, 131, 216-223.
- [10] Anderson, N. L., Copple, D. C., Bendele, R. A., Probst, G. S., Richardson, F. C., *Fundam. Appl. Toxicol.* 1992, 18, 570-580.

Progress with Proteome Projects: Why all Proteins Expressed by a Genome Should be Identified and How To Do It

MARC R. WILKINS¹, JEAN-CHARLES SANCHEZ¹, ANDREW A. GOOLEY¹,
RON D. APPEL¹, IAN HUMPHERY-SMITH², DENIS F. HOCHSTRASSER³
AND KEITH L. WILLIAMS^{1,*}

¹ Macquarie University Centre for Analytical Biotechnology, Macquarie University, Sydney, NSW 2109, Australia; ² Department of Microbiology, University of Sydney, NSW, 2006, Australia and ³ Central Clinical Chemistry Laboratory and Medical Computing Centre of the University of Geneva, CH 1211 Geneva 14, Switzerland

Introduction

The advent of large genome sequencing projects has changed the scale of biology. Over a relatively short period of time, we have witnessed the elucidation of the complete nucleotide sequence for bacteriophage λ (Sanger *et al.*, 1982), the nucleotide sequence of an eukaryotic chromosome (Oliver *et al.*, 1992), and in the near future will see the definition of all open reading frames of some simple organisms, including *Mycoplasma pneumoniae*, *Escherichia coli*, *Saccharomyces cerevisiae*, *Caenorhabditis elegans* and *Arabidopsis thaliana*. Nevertheless, genome sequencing projects are not an end in themselves. In fact, they only represent a starting point to understanding the function of an organism. A great challenge that biologists now face is how the co-expression of thousands of genes can best be examined under physiological and pathophysiological conditions, and how these patterns of expression define an organism.

There are two approaches that can be used to examine gene expression on a large scale. One uses nucleic acid-based technology, the other protein-based technology. The most promising nucleic-acid based technology is differential display of mRNA (Liang and Pardee, 1992; Bauer *et al.*, 1993), which uses polymerase chain reaction with arbitrary primers to generate thousands of cDNA species, each which correspond to an expressed gene or part of a gene. However, it is currently unclear if this technique can be developed to reliably assay the expression of thousands of genes or

* Corresponding Author

identify all cDNA species, and the approach does not easily allow a systematic screening. Analysis of gene expression by the study of proteins present in a cell or tissue presents a favorable alternative. This can be achieved by use of two-dimensional (2-D) gel electrophoresis, quantitative computer image analysis, and protein identification techniques to create 'reference maps' of all detectable proteins. Such reference maps establish patterns of normal and abnormal gene expression in the organism, and allow the examination of some post-translational protein modifications which are functionally important for many proteins. It is possible to screen proteins systematically from reference maps to establish their identities.

To define protein-based gene expression analysis, the concept of the 'proteome' was recently proposed (Wilkins *et al.*, 1995; Wasinger *et al.*, 1995). A proteome is the entire PROTEin complement expressed by a genome, or by a cell or tissue type. The concept of the proteome has some differences from that of the genome, as while there is only one definitive genome of an organism, the proteome is an entity which can change under different conditions, and can be dissimilar in different tissues of a single organism. A proteome nevertheless remains a direct product of a genome. Interestingly, the number of proteins in a proteome can exceed the number of genes present, as protein products expressed by alternative gene splicing or with different post-translational modifications are observed as separate molecules on a 2-D gel. As an extrapolation of the concept of the 'genome project', a 'proteome project' is research which seeks to identify and characterise the proteins present in a cell or tissue and define their patterns of expression.

Proteome projects present challenges of a similar magnitude to that of genome projects. Technically, the 2-D gel electrophoresis must be reproducible and of high resolution, allowing the separation and detection of the thousands of proteins in a cell. Low copy number proteins should be detectable. There should be computer gel image analysis systems that can qualitatively and quantitatively catalog the electrophoretically separated proteins, to form reference maps. A range of rapid and reliable techniques must be available for the identification and characterisation of proteins. As a consequence of a proteome project, protein databases must be assembled that contain reference information about proteins: such databases must be linked to genomic databases and protein reference maps. Databases should be widely accessible and easy to use.

Recently, there have been many changes in the techniques and resources available for the analysis of proteomes. It is the aim of this chapter to discuss the status of the areas outlined above, and to review briefly the progress of some current proteome projects.

Two-dimensional electrophoresis of proteomes

Two dimensional (2-D) gel electrophoresis involves the separation of proteins by their isoelectric point in the first dimension, then separation according to molecular weight by sodium dodecyl sulfate electrophoresis in the second dimension. Since first described (Klose, 1975; O'Farrell, 1975; Scheele, 1975), it has become the method of choice for the separation of complex mixtures of proteins, albeit with many modifications to the original techniques. 2-D electrophoresis forms the basis of proteome projects through separating proteins by their size and charge (Hochstrasser *et al.*,

HEPG2 2D-PAGE MAP

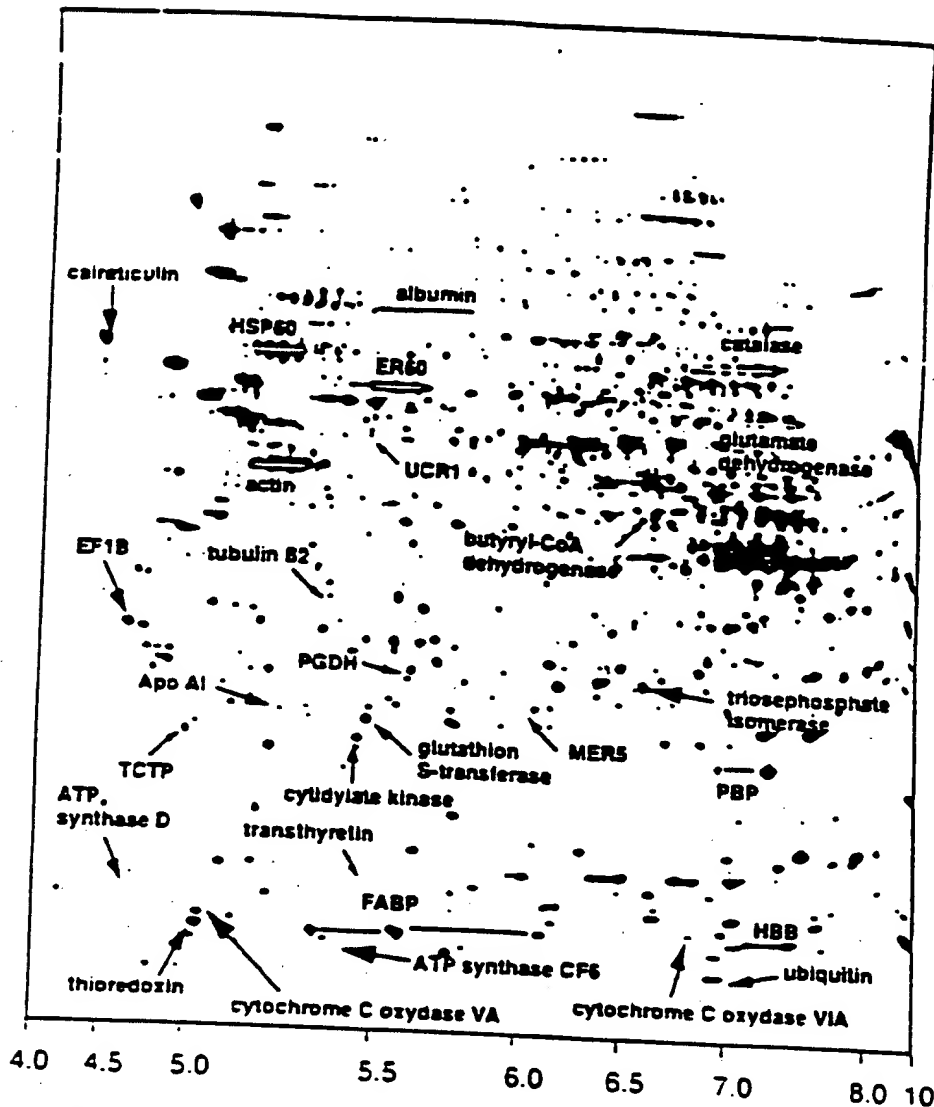


Figure 1. Two-dimensional gel electrophoresis map of a human hepatoblastoma-derived cell line, illustrating the very high resolution of the technique. The first dimensional separation (right to left of figure) was achieved using immobilised pH gradient electrophoresis of 4.0 to 10.0 units. The second dimension (top to bottom of figure) was SDS-PAGE using a 11%–14% acrylamide gradient, allowing separation in the molecular weight range 10–250 kDa. Proteins were visualised by silver staining. Arrows show proteins of known identity.

1992; Celis *et al.*, 1993; Garrels and Franza, 1989; VanBogelen *et al.*, 1992). Current protocols can resolve two to three thousand proteins from a complex sample on a single gel (Figure 1).

2-D GEL RESOLUTION AND REPRODUCIBILITY

A primary challenge of separating complex mixtures of proteins by 2-D gel electrophoresis has been to achieve high resolution and reproducibility. High resolution ensures that a maximum of protein species are separated, and high reproducibility is

vital to allow comparison of gels from day to day and between research sites. These factors can be difficult to achieve.

Carrier ampholytes are a common means of isoelectric focusing for the first dimension of 2-D electrophoresis. Gels are usually focused to equilibrium to separate proteins in the pI range 4 to 8, and run in a non-equilibrium mode (NEPHGE) to separate proteins of higher pI (7 to 11.5) (O'Farrell, 1975; O'Farrell, Goodman and O'Farrell, 1977). Unfortunately, the use of carrier ampholytes in the isoelectric focusing procedure is susceptible to 'cathode drift', whereby pH gradients established by prefocusing of ampholytes slowly change with time (Righetti and Drysdale, 1973). Carrier ampholyte pH gradients are also distorted by high salt concentration of samples (Bjellqvist *et al.*, 1982), and by high protein load (O'Farrell, 1975). A further limitation is that isoelectric focusing gels, which are cast and subject to electrophoresis in narrow glass tubes, need to be extruded by mechanical means before application to the second dimension – a procedure that potentially distorts the gel. Nevertheless, many of the above shortcomings can be avoided by loading small amounts of ^{14}C or ^{35}S radiolabelled samples (Garrels, 1989; Neidhardt *et al.*, 1989; Vandekerckhove *et al.*, 1990). High sensitivity detection is then achieved through use of fluorography or phosphorimaging plates (Bonner and Laskey, 1974; Johnston, Pickett and Barker, 1990; Patterson and Latter, 1993). However, this approach is only practicable for organisms or tissues that can be radiolabelled.

An alternative technique, which is becoming the method of choice for the first dimension separation of proteins, involves isoelectric focusing in immobilized pH gradient (IPG) gels (Bjellqvist *et al.*, 1982; Görg, Postel and Gunther, 1988; Righetti, 1990). Immobilized pH gradients are formed by the covalent coupling of the pH gradient into an acrylamide matrix, creating a gradient that is completely stable with time. IPG gels are usually poured onto a stiff backing film, which is mechanically strong and provides easy gel handling (Ostergren, Eriksson and Bjellqvist, 1988). The major advantages of IPG separations are that they do not suffer from cathodic drift, they allow focusing of basic and very acidic proteins to equilibrium, pH gradients can be precisely tailored (linear, stepwise, sigmoidal), and that separations over a very narrow pH range are possible (0.05 pH units per cm) (Righetti, 1990; Bjellqvist *et al.*, 1982, 1993a; Sinha *et al.*, 1990; Görg *et al.*, 1988; Gelfi *et al.*, 1987; Gunther *et al.*, 1988). However, it is not currently possible to use IPG gels to separate very basic proteins of isoelectric point greater than 10, although this is under development. Narrow pH range separations are useful to address problems of protein co-migration in complex samples, allowing 'zooming in' on regions of a gel (Figure 2). IPG gel strips are now commercially available, which begin to address the problems of intra- and inter-lab isoelectric focusing reproducibility.

There are two means of electrophoresis for the second dimension separation of proteins: vertical slab gels and horizontal ultrathin gels (Görg, Postel and Gunther, 1988). Both are usually SDS-containing gradient gels of approximately 11% to 15% acrylamide, which separate proteins in the molecular mass range of 10 – 150 kD. A stacking gel is not usually used with slab gels, but is necessary when using horizontal gel setups (Görg, Postel and Gunther, 1988). Comparisons have shown that there is little or no difference in the reproducibility of electrophoresis using either approach (Corbett *et al.*, 1994a), but commercially available vertical or horizontal precast gels will provide greater reproducibility for occasional users. For slab gel electrophoresis,

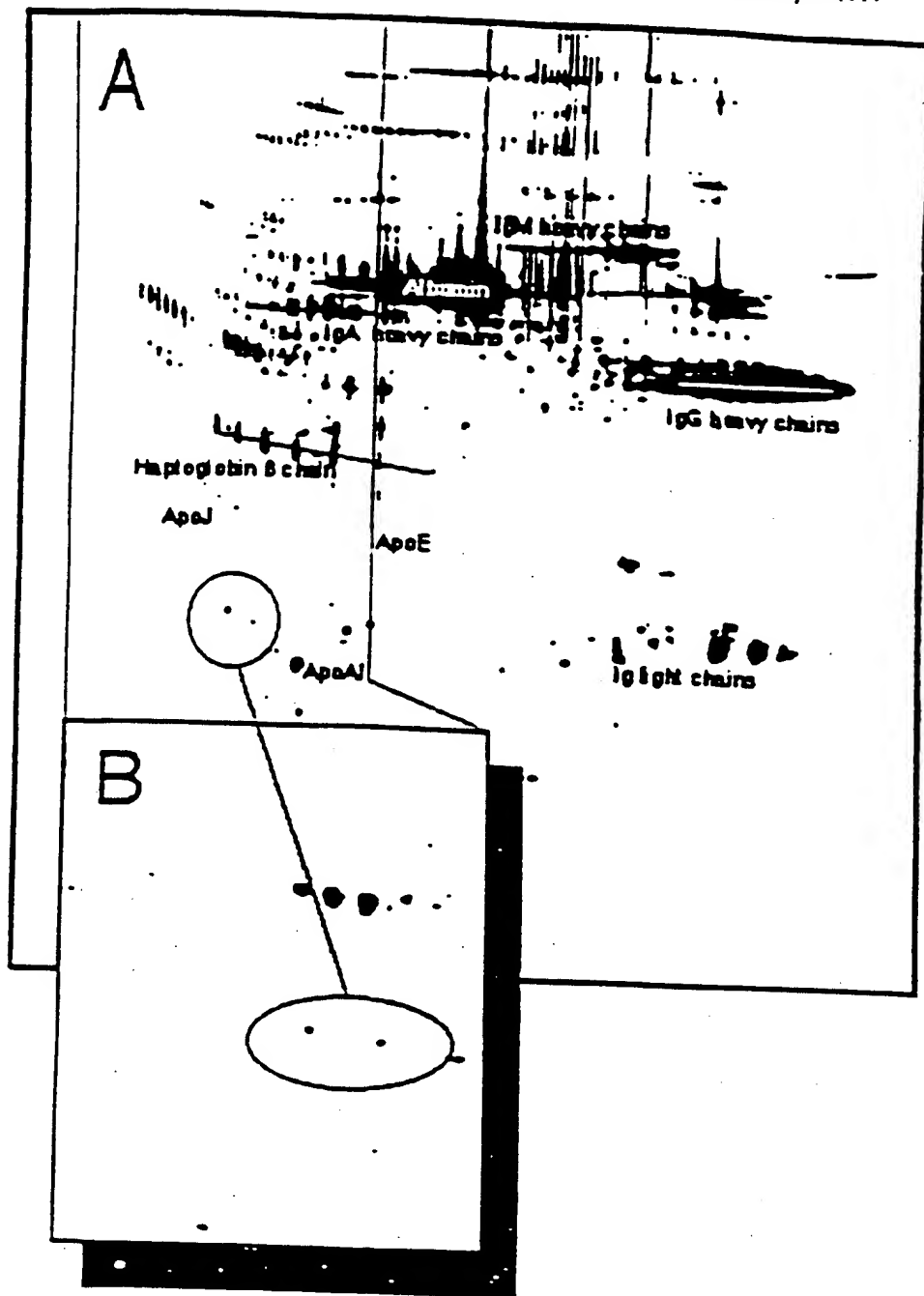


Figure 2. Two-dimensional gel electrophoresis allows 'zooming in' on areas of interest. Rings highlight 2 proteins common to each gel. (A) Wide pI range two dimensional electrophoresis map of human plasma proteins. First dimension separation was achieved using an immobilised pH gradient of 3.5 to 10.0 units. The second dimension was SDS-PAGE. Actual gel size was 16cm x 20cm, and proteins were visualised with silver staining. (B) Narrow pI range electrophoresis was used to 'zoom in' on a small region of the plasma map. The first dimension used a narrow range immobilised pH gradient of 4.2 to 5.2 units, and second dimension was SDS-PAGE. Micropreparative loading was used, and the gel blotted to PVDF. Proteins were visualised with amido black. Actual blot size was 16cm x 20cm.

the use of piperazine diacrylyl as a gel crosslinker and the addition of thiosulfate in the catalyst system has been shown to give better resolution and higher sensitivity detection (Hochstrasser and Merril, 1988; Hochstrasser, Patchornik and Merril, 1988).

Notwithstanding the advances described above, there is an increasing demand to improve the reproducibility of 2-D electrophoresis to facilitate database construction and proteome studies. Harrington *et al.* (1993) explain that if a gel resolves 4000 protein spots, and there is 99.5% spot matching from gel to gel, this will produce 20 spot errors per gel. This amount of error, which might accumulate with each gel to gel comparison used in database construction, could produce an unacceptable degree of uncertainty in gel databases. To address these issues, partial automation of large 2-D gel separations has been undertaken (Nokihara, Morita and Kuriki, 1992; Harrington *et al.*, 1993). Although results are preliminary, spot to spot positional reproducibility in one study was found to be threefold improved over manual methods (Harrington *et al.*, 1993). It should be noted that small 2-D gel formats (50 × 43 mm) have been almost completely automated (Brewer *et al.*, 1986), although these are not generally used for database studies.

MICROPREPARATIVE 2-D GEL ELECTROPHORESIS

With the advent of affordable protein microcharacterisation techniques, including N-terminal microsequencing, amino acid analysis, peptide mass fingerprinting, phosphate analysis and monosaccharide compositional analysis, a new challenge for 2-D electrophoresis has been to maintain high resolution and reproducibility but to provide protein in sufficient quantities for chemical analysis (high nanogram to low microgram quantities of proteins per spot). This becomes difficult to achieve with very complex samples such as whole bacterial cells, as the initial protein load is divided among 2000 to 4000 protein species. Two approaches are used for producing amounts of material that can be chemically characterised. The first method is to run multiple gels, collect and pool the spots of interest, and subject them to concentration (Ji *et al.*, 1994; Walsh *et al.*, 1995; Rasmussen *et al.*, 1992). In this approach, the concentration process must also act as a purification step to remove accumulated electrophoretic contaminants such as glycine. A more elegant approach has been to exploit the high loading capacity of IPG isoelectric focusing. The high loading capacity of immobilised pH gradients was described early (Ek, Bjellqvist and Righetti, 1983), but has only recently been applied to 2-D electrophoresis (Hanash *et al.*, 1991; Bjellqvist *et al.*, 1993b). Up to 15 mg of protein can be applied to a single gel, yielding microgram quantities of hundreds of protein species. A further benefit of this approach is that proteins present in low abundance, which may not be visualised by lower protein loads, are more likely to be detected. The use of electrophoretic or chromatographic prefractionation techniques (Hochstrasser *et al.*, 1991a; Harrington *et al.*, 1992), followed by high loading of narrow-range IPG separations (Bjellqvist *et al.*, 1993b) provides a likely solution to studies on proteins present in low abundance.

Methods of protein detection

There are many means for detecting proteins from 2-D gels. The method used will be dictated by factors including protein load on gel (analytical or preparative), the purpose of the gel (for protein quantitation or for blotting and chemical characterisation), and the sensitivity required. The most common means of protein detection and their applications are shown in Table 1. Most detection methods have drawbacks, for

Table 1: Common stains for 2-D gels or blots and their applications.

Detection Method	Main applications	Unsuitable applications	Sensitivity	References
[³⁵ S] Met or ¹⁴ C radiolabelling and fluorography or phosphorimaging	Cell lines, cultured organisms	Samples that cannot be labelled	20 ppm of radiolabel in a spot	Ganiels and Franza, 1988 Latham, Garret's and Soltis, 1993
[³⁵ S]thiourea silver	Extremely high sensitivity gel staining	Preparative 2-D, PVDF or NC membranes	0.4 ng protein on spot or band of gel	Wallace and Saluz, 1992a,b
Silver	Very high sensitivity gel staining, can be mono or polychromatic	Preparative 2-D, PVDF or NC membranes	4 ng protein on spot or band of gel	Rabilloud, 1992 Hochstrasser and Merril, 1988
Coomassie blue R-250	Staining of gels; staining of PVDF membranes before protein sequencing	Staining prior to direct mass determination from PVDF; amino acid analysis on PVDF; detection of some glycoproteins	30 ng protein on band or spot of gel	Sirupai <i>et al.</i> , 1994; Gharahdaghi <i>et al.</i> , 1992; Goldberg <i>et al.</i> , 1988; Sanchez <i>et al.</i> , 1992
Colloidal gold	Staining NC membranes, staining PVDF before direct MALDI-TOF	Gels	60 x higher than coomassie	Yamaguchi and Asakawa, 1988; Eckerskorn <i>et al.</i> , 1992; Sirupai <i>et al.</i> , 1994
Zinc imidazole	Reverse staining of gels or membranes; may be beneficial in MALDI-TOF of peptides	Where positive image is required	Higher than coomassie	Ortiz <i>et al.</i> , 1992; James <i>et al.</i> , 1993
Ponceau S and amido black	Staining higher protein loads on PVDF, for protein sequencing or amino acid analysis	Staining prior to direct mass determination from PVDF	100 ng protein on band or spot of gel	Sanchez <i>et al.</i> , 1992; Sirupai <i>et al.</i> , 1994; Wilkins <i>et al.</i> , 1995
India ink	Staining of membrane-bound proteins, staining PVDF before direct MALDI-TOF	Gel staining, not quantitative from protein to protein	1-10 ng	Li <i>et al.</i> , 1989; Hughes, Mack and Hambarian, 1988; Sirupai <i>et al.</i> , 1994
Stains-all	Staining to detect glycoproteins or Ca ²⁺ binding proteins	General gel staining	100 ng protein on band or spot of gel	Campbell, MacLennan and Jorgensen, 1983; Goldberg <i>et al.</i> , 1988

PVDF = polyvinylidene difluoride, NC = nitrocellulose, MALDI-TOF = matrix assisted laser desorption/ionisation time of flight mass spectrometry

example, some glycoproteins are not stained by coomassie blue (Goldberg *et al.*, 1988), and many organic dyes are unsuitable for protein detection on PVDF if samples are to be used for direct matrix-assisted laser desorption/ionisation mass spectrometry (Sirupai *et al.*, 1994).

Although most means of protein detection give some indication of the quantities of protein present, in general they cannot be used for global quantitation. This is because

no protein, stain is able consistently to detect proteins over a wide range of concentrations, isoelectric points and amino acid compositions, and with a variety of post-translational modifications (Goldberg *et al.*, 1988; Li *et al.*, 1989). Furthermore, there are large differences in staining pattern when identical gels or blots are subjected to different stains, including amido black, imidazole zinc, india ink, ponceau S, colloidal gold, or coomassie blue (Tovey, Ford and Baldo, 1987; Ortiz *et al.*, 1992). The most common means of quantitating large numbers of proteins in a 2-D gel involves the radiolabelling of protein samples prior to electrophoresis, and protein quantitation based on fluorography and image analysis or liquid scintillation counting (Garrels, 1989; Celis and Olsen, 1994). However, proteins which do not contain methionine cannot be detected if only [³⁵S] methionine is used for labelling. Amino acid analysis of protein spots visualised by other techniques presents a likely means of protein quantitation for the future.

BLOTTING OF PROTEINS TO MEMBRANES

Electrophoretic blotting of proteins from two-dimensional polyacrylamide gels to membranes presents many options for protein identification and microcharacterisation which are not possible when proteins remain in gels. For example, when proteins are blotted to polyvinylidene difluoride (PVDF) membranes, they can be identified by N-terminal sequencing, amino acid analysis, or immunoblotting, or they may be subjected to endoprotease digestion, monosaccharide analysis, phosphate analysis, or direct matrix-assisted laser desorption ionisation mass spectrometry (Matsudaira, 1987; Wilkins *et al.*, 1995; Jungblut *et al.*, 1994; Sutton *et al.*, 1995; Rasmussen *et al.*, 1994; Weizthandler *et al.*, 1993; Murthy and Iqbal, 1991; Eckerskorn *et al.*, 1992). It is possible to combine some of these procedures on a single protein spot on a PVDF membrane (Packer *et al.*, 1995; Wilkins *et al.*, submitted; Weizthandler *et al.*, 1993). This is useful when minimal amounts of protein are available for analysis. These techniques will be explored in detail later in this review. Notwithstanding the above, there are some disadvantages associated with blotting of proteins to membranes. There is always loss of sample during blotting procedures (Eckerskorn and Lottspeich, 1993), and common protein detection methods are less sensitive or not applicable to membranes (Table 1), presenting difficulties for the analysis of low abundance proteins. Detailed discussion of the merits of available membranes and common blotting techniques can be found elsewhere (Eckerskorn and Lottspeich, 1993; Strupat *et al.*, 1994; Patterson, 1994).

2-D gel analysis, documentation, and proteome databases

Following protein electrophoresis and detection, detailed analysis of gel images is undertaken with computer systems. For proteomic projects, the aim of this analysis is to catalogue all spots from the 2-D gel in a qualitative and if possible quantitative manner, so as to define the number of proteins present and their levels of expression. Reference gel images, constructed from one or more gels, form the basis of two-dimensional gel databases. These databases also contain protein spot identities and

details of their post-translational modifications. 2-D gel databases are beginning to be linked to or integrated with comprehensive protein and nucleic acid databases (Neidhardt *et al.*, 1989; Simpson *et al.*, 1992; Appel *et al.*, 1994), and 'organism' databases, containing DNA sequence data, chromosomal map locations, reference 2-D gels and protein functional information for an organism, are becoming established as genome and proteome projects progress (VanBogelen *et al.*, 1992; Yeast Protein Database cited in Garrels *et al.*, 1994).

GEL IMAGE ANALYSIS AND REFERENCE GELS

After 2-D electrophoresis and protein visualisation by staining, fluorography or phosphorimaging, images of gels are digitised for computer analysis by an image scanner, laser densitometer, or charge-coupled device (CCD) camera (Garrels, 1989; Celis *et al.*, 1990a; Urwin and Jackson, 1993). All systems digitise gels with a resolution of 100 – 200 µm, and can detect a wide range of densities or shading (256 or more 'grey scales'). Following this, gel images are subjected to a series of manipulations to remove vertical and horizontal streaking and background haze, to detect spot positions and boundaries, and to calculate spot intensity (Figure 3). A standard spot (SSP) number, containing vertical and horizontal positional information, is assigned to each detected spot and becomes the protein's reference number. Table 2 lists some notable software packages which process 2-D gel images.

Table 2: Some Software Packages for the Analysis of Gel Images.

Gel Image Analysis System	References*
ELSIE I & II	Olsen and Miller, 1988; Wirth <i>et al.</i> , 1991; Wirth <i>et al.</i> , 1993
GELLAB I & II	Wu, Lemkin and Upton, 1993; Lemkin, Wu and Upton, 1993; Myrick <i>et al.</i> , 1993
MELANIE I & II	Appel <i>et al.</i> , 1991; Hochstrasser <i>et al.</i> , 1991b
QUEST I & II and PDQUEST	Garrels, 1989; Monardo <i>et al.</i> , 1994; Holt <i>et al.</i> , 1992; Celis <i>et al.</i> , 1990a,b
TYCHO & KEPLAR	Anderson <i>et al.</i> , 1984; Richardson, Horn and Anderson, 1994

* These references are not exhaustive; they include some references of use as well as authors of the system.

As there are difficulties in the electrophoresis of samples with 100% reproducibility, reference gel images are often constructed from many gels of the same sample (Garrels and Franza, 1989; Neidhardt *et al.*, 1989). Since this involves the matching of 2000 to 4000 proteins from one gel to another, it presents a considerable challenge to image analysis systems. Matching of gels is usually initiated by an operator, who manually designates approximately 50 or so prominent spots as 'landmarks' on gels to be cross-matched. Proteins which match are then established around landmarks, using computer-based vector algorithms to extend the matching over the entire gel. Close to 100% of spots from complex samples can be matched by these methods, although different degrees of operator intervention may be required (Olsen and Miller, 1988; Lemkin and Lester, 1989; Garrels, 1989; Myrick *et al.*, 1993).

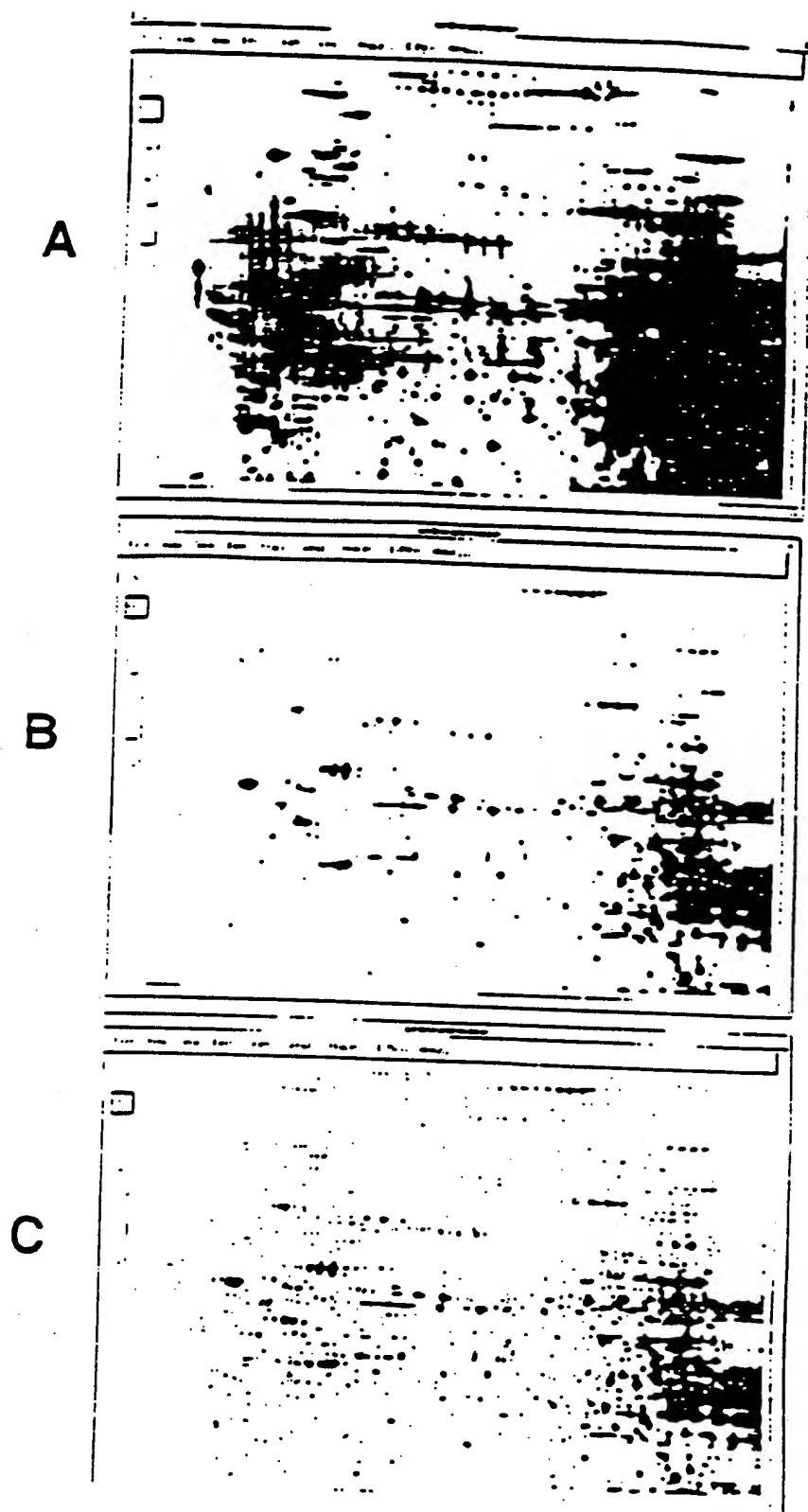


Figure 3. Computer processing of gel images. Shown is a wide pI range 2-D separation of human liver proteins, processed by Melanie software (Appel *et al.*, 1991). (A) Original gel image as captured by laser densitometer. (B) Gel image after processing to remove streaking and background. (C) Outline definition of all spots on the gel.

CALCULATION OF PROTEIN ISOELECTRIC POINT AND MOLECULAR WEIGHT

Estimation of the isoelectric point (pI) and molecular weight (MW) of proteins from 2-D gels provides fundamental parameters for each protein, which are also of use during identification procedures (see following section). The pI and MW of proteins are recorded in 2-D gel databases. Accurate estimations of protein pI and MW can be obtained by using 20 or more known proteins on a reference map to construct standard curves of pI and molecular weight, which are then used to calculate estimated pI and MW of unknown proteins (Neidhardt *et al.*, 1989; Garrels and Franza, 1989; Van-Bogelen, Hutton and Neidhardt, 1990; Anderson and Anderson, 1991; Anderson *et al.*, 1991; Latham *et al.*, 1992). Alternatively, the MW of individual proteins blotted to PVDF can be determined very accurately by direct mass spectrometry (Eckerskorn *et al.*, 1992). Where immobilised pH gradients are used, the focusing position of proteins allows their pI to be measured within 0.15 units of that calculated from the amino acid sequence (Bjellqvist *et al.*, 1993c). It must be noted, however, that proteins carrying post-translational modifications may migrate to unexpected pI or MW positions during electrophoresis (Packer *et al.*, 1995).

SPOT QUANTITATION AND EXPRESSION ANALYSIS

A major challenge faced in proteome projects is the quantitative analysis of proteins separated by 2-D electrophoresis. The most accurate means of protein quantitation is to determine chemically the amount of each protein present by amino acid compositional analysis. However, the current method of choice for quantitative analysis of many proteins is to radiolabel samples with [³⁵S] methionine or ¹⁴C amino acids, perform the 2-D electrophoresis, and measure protein levels in disintegrations per minute (dpm) or units of optical density. Quantitation is achieved either by liquid scintillation counting, or by gel image analysis where spot densities are quantitated by reference to gel calibration strips containing known amounts of radiolabelled protein or against the integrated optical density of all spots visualised (Vandekerckhove *et al.*, 1990; Celis *et al.*, 1990b; Celis and Olsen, 1994; Garrels, 1989; Latham, Garrels and Solter, 1993; Fey *et al.*, 1994). All approaches effectively allow spots to be normalised against the total disintegrations per minute loaded onto the gel. Limitations that remain with radiolabelling methods are that absolute quantitation is not achieved because all proteins have varying amounts of any amino acid, and that only easily labelled samples can be investigated. Quantitative silver staining presents an alternative (Giometti *et al.*, 1991; Harrington *et al.*, 1992; Rodriguez *et al.*, 1993; Myrick *et al.*, 1993), which when undertaken with [³⁵S]thiourea (Wallace and Saluz, 1992 a,b) is of extremely high sensitivity.

When protein spots from samples prepared under different conditions are quantitated and matched from gel to gel, it becomes possible to examine changes and patterns in protein expression. Large scale investigation of up- and down-regulation of proteins, their appearance and disappearance, can be undertaken. For example, simian virus 40 transformed human keratinocytes were shown to have 177 up-regulated and 58 down-regulated proteins compared to normal keratinocytes (Celis and Olsen, 1994); detailed synthesis profiles of 1200 proteins have been established in 1 to 4 cell mouse embryos (Latham *et al.*, 1991, 1992); and 4 proteins out of 1971 were found to be markers for

cadmium toxicity in urinary proteins (Myrick *et al.*, 1993). Complex global changes in protein expression as a result of gene disruptions have also been investigated (S. Fey and P. Moss-Larsen, Personal communication). Impressively, large gel sets showing protein expression under different conditions can be globally investigated using statistical methods that find groups of related objects within a set. For example, the REF52 rat cell line database, consisting of 79 gels from 12 experimental groups where each gel contains quantitative data for 1600 cross-matched proteins, has been analysed by cluster analysis (Garrels *et al.*, 1990). This revealed clusters of proteins that, for example, were induced or repressed similarly under simian virus 40 or adenovirus transformation, suggesting a common mechanism. Protein groups that were induced or repressed during culture growth to confluence were also found. It is obvious that the potential for investigation of cellular control mechanisms by these approaches is immense. It is equally clear that investigations of gene expression of this scale are currently technically impossible using nucleic-acid based techniques.

Table 3: Some proteomic databases and their special features

Proteomic database	Special features	References
<i>E. coli</i> gene-protein database	Gel spots linked with GenBank and Kohara clones; quantitative spot measurements under different growth conditions	VanBogelen and Neidhardt, 1991. VanBogelen <i>et al.</i> , 1992
Human heart databases	Identification of disease markers; two separate databases have been established	Baker <i>et al.</i> , 1992 Corbett <i>et al.</i> , 1994b Jungblut <i>et al.</i> , 1994
Human keratinocyte database	Extensive identifications; quantitative spot measurements of transformed cells; identification of disease markers	Celis <i>et al.</i> , 1990a Celis <i>et al.</i> , 1993 Celis and Olsen, 1994
Mouse embryo database	Quantitative spot measurements through 1 to 4 cell stage	Latham <i>et al.</i> , 1991 Latham <i>et al.</i> , 1992
Mouse liver database (Argonne Protein Mapping Group)	Documents changes due to exposure to ionizing radiation and toxic chemicals	Gonnert, Taylor and Tollaksen, 1992
Rat liver epithelial database	Detailed subcellular fractionation studies	Wirth <i>et al.</i> , 1991; Wirth <i>et al.</i> , 1993
Rat liver database	Extensive studies on regulation of proteins by drugs and toxic agents	Anderson and Anderson, 1991. Anderson <i>et al.</i> , 1992; Richardson, Horn and Anderson, 1994
REF 52 rat cell line database	Accessible via World Wide Web; quantitative spot measurements under different conditions	Garrels and Franza 1989 Boutell <i>et al.</i> , 1994
SWISS-2DPAGE containing human reference maps	Accessible via World Wide Web; completely integrated with SWISS-PROT and SWISS-3DIMAGE	Appel <i>et al.</i> , 1993 Hochstrasser <i>et al.</i> , 1992 Hughes <i>et al.</i> , 1993 Golar <i>et al.</i> , 1993
Yeast Protein Database (YPD) and Yeast Electrophoretic Protein Database (YEPD)	Completely crossreferenced organism database; YPD has extensive information on over 3500 proteins; YEPD has many identifications	Garrels <i>et al.</i> , 1994

Prote
prote
infer
2-D g
subse
of rel
shoul
Macr
the ar
annot
seque
On
SWIS
1994;
featur
2DPA

Table 4
All thre
expans

Informa

Annua

Cross-
Refer
Datah.

Other

FEATURES OF PROTEOME DATABASES

Proteome projects rely heavily on computer databases to store information about all proteins expressed by an organism. 'Proteome databases' should contain detailed information of proteins already characterised elsewhere, as well as protein data from 2-D gels such as apparent pI and MW, expression level under different conditions, subcellular localisation, and information on post-translational modifications. Images of reference 2-D gels, showing protein SSP numbers and protein identifications, should also be included. Ideally, proteome databases should be accessible with Macintosh or IBM personal computers and easy to use. Some proteome databases and the areas they cover are listed in *Table 3*. Databases range from collections of unannotated gels to large databases of images integrated with protein and nucleic acid sequence banks.

One example of an integrated proteome database is the suite of SWISS-PROT, SWISS-2DPAGE and SWISS-3DIMAGE databases (Appel *et al.*, 1993; Appel *et al.*, 1994; Appel, Bairoch and Hochstrasser, 1994; Bairoch and Boeckmann, 1994). The features of these three databases are listed in *Table 4*. SWISS-PROT, SWISS-2DPAGE and SWISS-3DIMAGE are accessible through the World Wide Web

Table 4: The SWISS-PROT, SWISS-2DPAGE and SWISS-3DIMAGE suite of crosslinked databases. All three databases are accessible through the World Wide Web, at URL address: <http://expasy.hugboch.ch/>

	SWISS-PROT	SWISS-2DPAGE	SWISS-3DIMAGE
Information	Text entries of sequence data; Citation information; taxonomic data, 38, 303 entries in Release 29	2-D gel images of: human liver, plasma, HepG2, HepG2 secreted proteins, red blood cell, lymphoma, cerebrospinal fluid, macrophage like cell line, erythroleukemia cell, platelet	Collection of 330 3-D images of proteins
Annotations	Protein function, Post translational modifications, Domains, Secondary structure, Quaternary structure, Diseases associated with protein, Sequence conflicts	Gel images where protein is found; How protein identified, Protein pI and MW, protein number, normal and pathological variants	All annotation is available in SWISS- PROT
Cross- Referenced Databases	SWISS-2DPAGE SWISS-3DIMAGE EMBL, PIR, PDB, ONIM, PROSITE, Medline, Flybase, GCRDB, MaizeDB, WonnPep, DictyDB	SWISS-PROT and all other databases accessible through SWISS-PROT	SWISS-PROT and all other databases accessible through SWISS-PROT
Other Features	Navigation in other SWISS databases achieved by selecting entries with computer mouse	Gel images show position of identified proteins, or region of gel where protein should appear	Mono and stereo images available, Images can be transferred to local computer image viewing programs

(Berners-Lee *et al.*, 1992), allowing any computer connected to the internet to access the stored information and images. Navigation within and between the three databases is seamless, as all potential crosslinks are highlighted as hypertext on the display and can be selected with a computer mouse. From these databases, detailed information about a protein, including amino acid sequence and known post-translational modifications, can be obtained, the precise protein spot it corresponds to on a reference gel image can be viewed if known, and the 3-D structure of the molecule can be seen if available. References to nucleic acid and other databases are also given to provide access to information stored elsewhere.

Organism databases, containing detailed protein and nucleic acid information about a species, are becoming common as genome and proteome projects progress. These differ from nucleic acid or protein sequence databases like GenBank or SWISS-PROT because they are image based, and contain information about chromosomal map positions, transcription of genes, and protein expression patterns. The *Escherichia coli* gene-protein database (VanBogelen, Hutton and Neidhardt, 1990; VanBogelen and Neidhardt, 1991; VanBogelen *et al.*, 1992), known as the ECO2DBASE, is one example. It contains gene and protein names, 2-D gel spot information (including pI and MW estimates, and spot identification), genetic information (GenBank or EMBL codes, chromosomal location, location on Kohara clones (Kohara, Akiyama, and Isono, 1987), transcription direction of genes), and protein regulatory information (level of protein expression under different growth regimes, member of regulon or sigma factor). All entries in the ECO2DBASE are also cross-referenced to the SWISS-PROT database (Bairoch and Boeckmann, 1994). It is anticipated that organism databases will soon become a standard means of storing all available information about a particular species. However there is currently no consistent manner in which organism databases are assembled, which may hamper comparisons in the future.

Identification and characterisation of proteins from 2-D gels

The number of proteins identified on a 2-D reference map determines its usefulness as a research and reference tool. As most reference maps have only a small proportion of proteins identified, a major aim of current proteome projects is to screen many proteins from 2-D maps, in order to define them as 'known' in current nucleic acid and protein databases, or as 'unknown'. Protein identification assists in confirmation of DNA open reading frames, and provides focus for DNA sequencing projects and protein characterisation efforts by pointing to proteins that are novel. Since there may be 3000-10000 proteins from a single 2-D map that require identification, the challenge in protein screening is to identify proteins quickly, with a minimum of cost and effort.

Traditionally, proteins from 2-D gels have been identified by techniques such as immunoblotting, N-terminal microsequencing, internal peptide sequencing, comigration of unknown proteins with known proteins, or by overexpression of homologous genes of interest in the organism under study (Matsudaira, 1987; Rosenfeld *et al.*, 1992; VanBogelen *et al.*, 1992; Celis *et al.*, 1993; Honore *et al.*, 1993; Garrels *et al.*, 1994). Whilst these techniques are powerful identification tools, they are too expensive or time and labour intensive to use in mass screening programs. A hierarchical approach to mass protein identification has been recently suggested as an

Table 5: Hierarchical analysis for mass screening of 2-D separated proteins blotted to membranes. Rapid and inexpensive techniques are used as a first step in protein identification, and slower, more expensive techniques are then used if necessary. Table modified from Wasinger *et al.*, 1995.

Order	Identification technique	References
1	Amino acid analysis	Jungblut <i>et al.</i> , 1992; Shaw, 1993; Hohmann, Houthaeve and Sander, 1993; Jungblut <i>et al.</i> , 1994; Wilkins <i>et al.</i> , 1995
2	Amino acid analysis with N-terminal sequence tag	Wilkins <i>et al.</i> , submitted
3	Peptide-mass fingerprinting	Henzel <i>et al.</i> , 1993; Pappin, Hourup and Bleasby, 1993; James <i>et al.</i> , 1993; Mann, Hourup and Roepstorff, 1993; Yates <i>et al.</i> , 1993; Moritz <i>et al.</i> , 1992; Sutton <i>et al.</i> , 1995
4	Combination of amino acid analysis and peptide mass fingerprinting	Cordwell <i>et al.</i> , 1995; Wasinger <i>et al.</i> , 1995
5	Mass spectrometry sequence tag	Mann and Wilm, 1994
6	Extensive N-terminal Edman microsequencing	Matsudaira, 1987
7	Internal peptide Edman microsequencing	Rosenfeld <i>et al.</i> , 1992; Hellman <i>et al.</i> , 1995
8	Microsequencing by mass spectrometry (electrospray ionisation, post-source decay MALDI-TOF)	Johnson and Walsh, 1992
9	Ladder sequencing	Barilei-Jones <i>et al.</i> , 1993

alternative to traditional approaches (Table 5; Wasinger *et al.*, 1995). This involves the use of rapid and cheap identification tools such as amino acid analysis and peptide mass fingerprinting as first steps in protein identification, followed by the use of slower, more expensive and time consuming identification procedures if necessary. In the construction of this hierarchy the analysis time, cost per sample and the complexity of the data created has been considered, as whilst some techniques require little machine time per sample, the analysis of data can be quite involved and time consuming. Amino acid analysis and peptide mass-fingerprinting based identification techniques in the hierarchy are discussed in detail below. For review of other protein identification techniques in Table 5, see Patterson (1994) and Mann (1995).

PROTEIN IDENTIFICATION BY AMINO ACID COMPOSITION

There has been a revival of interest in the use of amino acid composition for identification of proteins from 2-D gels after early work by Eckerskorn *et al.* (1988). This technique uses a protein's idiosyncratic amino acid composition profile in order to identify it by comparison with theoretical compositions of proteins in databases. The amino acid composition of proteins can be determined by differential metabolic radiolabelling and quantitative autoradiography after 2-D electrophoresis (Garrels *et al.*, 1994; Frey *et al.*, 1994), or by acid hydrolysis of membrane-blotted proteins and chromatographic analysis of the resulting amino acid mixture (Eckerskorn *et al.*, 1988; Tous *et al.*, 1989; Gharahdaghi *et al.*, 1992; Jungblut *et al.*, 1992; Wilkins *et al.*, 1995). As differential metabolic labelling experiments require X-ray film or phosphor-image plate exposures of up to 140 days, and can only be undertaken with easily radiolabelled samples, the technique is not as rapid or widely applicable as chromato-

Spot: ECOLI-BLM

Composition:

Asx: 13.2 Glx: 10.4 Ser: 5.7 His: 0.7
 Gly: 5.4 Thr: 3.6 Ala: 6.7 Pro: 7.9
 Tyr: 1.3 Arg: 5.0 Val: 8.0 Met: 0.3
 Ile: 5.9 Leu: 8.0 Phe: 13.3 Lys: 4.4

pI estimate: 6.89 Range searched: (6.64, 7.14)

Mw estimate: 16200 Range searched: (13440, 20160)

Closest SWISS-PROT entries for the species ECOLI matched by AA composition:

Rank	Score	Protein	pI	Mw	Description
1	24	PYRI_ECOLI	6.84	16989	ASPARTATE CARBAMOYLTRANSFERASE
2	39	CDAA_ECOLI	6.32	36359	PANTOTHENATE KINASE (EC 2.7.1.33)
3	40	META_ECOLI	5.06	35713	HOMOSERINE O-SUCCINYLTRANSFERASE
4	42	CADC_ECOLI	5.52	57812	TRANSCRIPTIONAL ACTIVATOR CADC.
5	43	HLYS_ECOLI	8.38	19769	HEMOLYSIN C. PLASMID.

Closest SWISS-PROT entries for ECOLI with pI and Mw values in specified range:

Rank	Score	Protein	pI	Mw	Description
1	24	PYRI_ECOLI	6.84	16989	ASPARTATE CARBAMOYLTRANSFERASE
2	102	TRAJ_ECOLI	6.73	17921	TRAJ PROTEIN.
3	112	YAJG_ECOLI	6.79	19028	HYPOTHETICAL LIPOPROTEIN YAJG.
4	140	YFJB_ECOLI	6.83	14945	HYPOTHETICAL 14.9 KD PROTEIN IN GRPE
5	142	YAHK_ECOLI	7.06	14726	HYPOTHETICAL PROTEIN IN BETT 3' REGION

Figure 4. Computer printout from ExPASy server where the empirical amino acid composition, estimated pI and MW of a protein from a 2-D reference map of *E. coli* were matched against all entries in SWISS-PROT for *E. coli*. The correct identification, aspartate carbamoyltransferase, is shown in bold. Low scores indicate a good match. Note how matching within a defined pI and MW range (lower set of proteins) has greatly increased the score difference between the first and second ranking proteins. This score difference gives high confidence in the identification, and is only observed where the top ranking protein is the correct identification (Wilkins *et al.*, 1995).

graphy-based analysis. Proteins blotted to PVDF membranes can be hydrolysed in 1 h at 155°C, amino acids extracted in a single brief step, and each sample automatically derivatised and separated by chromatography in under 40 minutes (Wilkins *et al.*, 1995; Ou *et al.*, 1995). In this manner, one operator can routinely analyse 100 proteins per week on one HPLC unit. This technology lends itself to automation, and it is anticipated that instruments with even greater sample throughput will be developed. When proteins have been prepared by micropreparative 2-D electrophoresis (Hanash *et al.*, 1991; Bjellqvist *et al.*, 1993b), blotted to a PVDF membrane and stained with amido black, any visible protein spot is of sufficient quantity for amino acid analysis (Cordwell *et al.*, 1995; Wasinger *et al.*, 1995; Wilkins *et al.*, 1995).

After the amino acid composition of a protein has been determined, computer programs are used to match it against the calculated compositions of proteins in databases (Eckerskorn *et al.*, 1988; Sibbald, Sommerfeldt and Argos, 1991; Jungblut *et al.*, 1992; Shaw, 1993; Hobohm, Houthaeve and Sander, 1994; Wilkins *et al.*, 1995). Matching is usually done with only 15 or 16 amino acids, as cysteine and

Figure
same
and
PRO
for it
large
the c
print

try
to it
The
a se
rest
199
et al
mat
in F
refe
ram
lym
et al

PRC
SEQ
Wh

Spot: ECOLI-ACC

Composition:

Asx: 5.6 Glx: 10.8 Ser: 4.1 His: 2.7
 Gly: 12.2 Thr: 3.8 Ala: 11.9 Pro: 3.2
 Tyr: 6.0 Arg: 3.7 Val: 9.5 Met: 0.6
 Ile: 5.0 Leu: 8.2 Phe: 3.2 Lys: 4.9

pI estimate: 5.99 Range searched: (5.74, 6.24)

Mw estimate: 45000 Range searched: (36000, 54000)

Closest SWISS-PROT entries for ECOLI with pI and Mw values in specified range:

Rank	Score	Protein	pI	Mw	N-terminal Seq.
1	21	GLYA_ECOLI	6.03	45316	M L K R E
2	32	YGGH_ECOLI	5.86	36502	M S H I K
3	38	GABT_ECOLI	5.78	45774	M S N S K
4	44	YTHS_ECOLI	5.86	48018	M R I K Y
5	45	DHE4_ECOLI	5.98	48581	M D Q T Y
6	46	ARSD_ECOLI	5.79	43765	M A I E Q
7	46	MYPB_ECOLI	5.78	37851	M N H S L
8	47	GLMT_ECOLI	5.98	49162	M L N R A
9	47	ACKA_ECOLI	5.85	43290	M S S K L
10	50	YGGH_ECOLI	6.01	37064	M E S R I

Figure 5. A PVDF protein spot from an *E. coli* 2-D reference map was sequenced for 4 cycles, and the same sample then subject to amino acid analysis. The N-terminal sequence was M L K R. When the amino acid composition of the spot, as well as estimated pI and MW, were matched against all entries in SWISS-PROT for *E. coli*, the above list of best matches was produced. N-terminal sequences are from SWISS-PROT for those entries. The top ranking identification of serine hydroxymethyltransferase (bold) did not show a large score difference between the first and second ranking proteins, giving little confidence in this being the correct protein identification. However, the sequence tag (M L K R) confirmed the identity of the protein as serine hydroxymethyltransferase.

tryptophan are destroyed during hydrolysis, asparagine and glutamine are deamidated to their corresponding acids, and proline is not quantitated in some analysis systems. The computer programs produce a list of best matching proteins, which are ranked by a score that indicates the match quality. Some programs allow matching to be restricted to specific 'windows' of MW and pI (Hobohm, Houthaeve and Sander, 1994; Wilkins *et al.*, 1995), and to protein database entries for one species (Jungblut *et al.*, 1992; Wilkins *et al.*, 1995). The use of such restrictions increases the power of matching. An example of protein identification by amino acid composition is shown in Figure 4. To date, amino acid composition has been used to identify proteins from reference maps of *Spiroplasma melliferum*, *Mycoplasma genitalium*, *E. coli*, *Saccharomyces cerevisiae*, *Dictyostelium discoideum*, human sera, human heart, human lymphocyte, and mouse brain (Cordwell *et al.*, 1995; Wasinger *et al.*, 1995; Wilkins *et al.*, 1995; Jungblut *et al.*, 1992, 1994; Garrels *et al.*, 1994; Frey *et al.*, 1994).

PROTEIN IDENTIFICATION BY AMINO ACID COMPOSITION AND N-TERMINAL SEQUENCE TAG

When samples from 2-D gels are not unambiguously identified by amino acid

composition, pI and MW, often the correct identification of that protein is amongst the top rankings of the list (Hobohm, Houthaeve and Sunder, 1994; Cordwell *et al.*, 1995; Wilkins *et al.*, 1995). Taking advantage of this observation, we have used the mass spectrometry 'sequence tag' concept (Mann and Wilkins, 1994) in developing a combined Edman degradation and amino acid analysis approach to protein identification (Wilkins *et al.*, submitted). This involves the N-terminal sequencing of PVDF-blotted proteins by Edman degradation for 3 or 4 cycles to create a 'sequence tag', following which the same sample is used for amino acid analysis. As only a few amino acids are removed from the protein, its composition is not significantly altered. Furthermore, since only a small amount of protein sequence is required, fast but low repetitive yield Edman degradation cycles can be used. Modifications to current procedures should allow 3 cycles to be completed in 1 h, thereby allowing the screening of 100 or more proteins per week on one automated, multi-cartridge sequenator. Amino acid composition, pI and MW of proteins are matched against databases as described above, and N-terminal sequences of best matching proteins are checked with the 'sequence tag' to confirm the protein identity (Figure 5). This technique will be less useful when proteins are N-terminally blocked, but as only a few N-terminal amino acids are susceptible to the acetyl, formyl, or pyroglutamyl modifications that cause blockage, this may itself provide useful information for sequence tag identification. A strength of N-terminal sequence tag and amino acid composition protein identification is that data generated are quickly and easily interpreted.

PROTEIN IDENTIFICATION BY PEPTIDE MASS FINGERPRINTING

Techniques for the identification of proteins by peptide mass fingerprinting have recently been described (Henzel *et al.*, 1993; Pappin, Hojrup and Bleasby, 1993; James *et al.*, 1993; Mann, Hojrup and Roepstorff, 1993; Yates *et al.*, 1993; Mortz *et al.*, 1994; Sutton *et al.*, 1995). This involves the generation of peptides from proteins using residue-specific enzymes, the determination of peptide masses, and the matching of these masses against theoretical peptide libraries generated from protein sequence databases. As proteins have different amino acid sequences, their peptides should produce characteristic 'fingerprints'.

The first step of peptide mass fingerprinting is protein digestion. Proteins within the gel matrix or bound to PVDF can be enzymatically digested *in situ*, although *in situ* gel digests are reported to produce more enzyme autodigestion products, which complicate subsequent peptide mass analysis (James *et al.*, 1993; Rasmussen *et al.*, 1994; Mortz *et al.*, 1994). The enzyme of choice for digestion is currently trypsin (of modified sequencing grade), but other enzymes (Lys-C or *S. aureus* V8 protease) have also been used (Pappin, Hojrup and Bleasby, 1993). To maximise the number of peptides obtained, it is desirable for protein samples to be reduced and alkylated prior to digestion (Mortz *et al.*, 1994; Henzel *et al.*, 1993). This ensures that all disulfide bonds of the protein are broken, and produces protein conformations that are more amenable to digestion. Surprisingly, chemical digestion methods such as cyanogen bromide (methionine specific), formic acid (aspartic acid specific), and 2-(2'-nitrophenyl)sulfonyl-3-methyl-3'-bromoindolenine (tryptophan specific) have not been explored as means of peptide production for mass fingerprinting, even though they are rapid and may circumvent some problems associated with enzyme digestions.

(Nikodem and Fresco, 1979; Crimmins *et al.*, 1990; Vanfleteren *et al.*, 1992).

After proteins are digested, peptide masses are determined by mass spectrometry. Direct analysis of peptide mixtures can be achieved by electrospray ionisation mass spectrometry, plasma desorption mass spectrometry, or matrix assisted laser desorption ionization (MALDI) mass spectrometry techniques. MALDI is preferable because of its higher sensitivity and greater tolerance to contaminating substances from 2-D gels (James *et al.*, 1993; Mortz *et al.*, 1994; Pappin, Hojrup and Bleasby, 1993). Furthermore, recent modifications to sample preparation methods have largely solved early difficulties experienced with the calibration of MALDI spectra (Mortz *et al.*, 1994; Vorm and Mann, 1994; Vorm, Roepstorff and Mann, 1994). The high sensitivity of mass spectrometry allows a small fraction of a digest of a 1 µg protein spot to be used for analysis, and analysis itself is complete in a few minutes.

A major challenge associated with peptide mass fingerprinting is data interpretation prior to computer matching against libraries of theoretical peptide digests. Spectra must be examined carefully to determine which peaks represent peptide masses of interest, as there are often enzyme autodigestion products and contaminating substances present (Henzel *et al.*, 1993; Mortz *et al.*, 1994; Rasmussen *et al.*, 1994). Furthermore, if protein alkylation and reduction has not been undertaken prior to protein digestion, peptide sequence coverage may be poor (40% to 70%), with some masses present representing disulfide bonded peptides originally present in the protein (Mortz *et al.*, 1994). For eukaryotes, a serious issue is the alteration of peptide masses by the presence of post-translational modifications (Table 6). The mass of the unmodified peptide alone can be very difficult to determine. Two artifactual modifications introduced by electrophoresis, an acrylamide adduct to cysteine and the oxidation of methionine, are also known to alter peptide masses (Le Maire *et al.*, 1993; Hess *et al.*, 1993).

Table 6: Masses of some common post-translational modifications. Peptides carrying post-translational modifications complicate data analysis for peptide mass fingerprinting protein identification. This is especially so for protein glycosylation, which involves many different combinations of the hexosamines, hexoses, deoxyhexoses, and sialic acid

Post-translational modification	Mass change
Acetylation	- 42.04
* Acrylamide adduct to cysteine	- 71.00
Carboxylation of Asp or Glu	- 71.00
Deamidation of Asn or Gln	- 24.01
Disulfide bond formation	- 0.98
Deoxyhexoses (Fuc)	- 2.02
Formylation	126.14
Hexosamines (GlcN, GalN)	- 28.01
Hexoses (Glc, Gal, Man)	- 161.16
Hydroxylation	- 162.14
N-acetylhexosamines (GlcNAc, GalNAc)	- 16.00
* Oxidation of Met	- 203.19
Phosphorylation	- 16.00
Pyroglutamic acid formed from Gln	- 79.98
Sialic acid (NeuNAc)	- 17.03
Sulfation	- 291.26
	- 80.06

Table modified from Finnigan LASERBIAT application data sheet 5

Asterisk * shows modifications that can arise artifactually from the 2-D electrophoresis process

A number of computer programs are available for matching peptide masses against databases (reviewed in Cottrell, 1994). Matching is usually undertaken in an interactive manner, whereby peaks of mass 500–3000 Da are selected and matched under various search parameters including MW of protein, mass accuracy of peptides, and number of missed enzyme cleavages allowed (Henzel *et al.*, 1993; Moritz *et al.*, 1994; Rasmussen *et al.*, 1994). The correct protein identity is the protein which has the most peptide masses in common with the unknown sample. Identities have been established with as few as three peptides, but unambiguous identification is thought to require a mass spectrometric map covering most peptides of the protein (Moritz *et al.*, 1994; Yates *et al.*, 1993). To date, peptide mass fingerprinting of proteins has been undertaken from the human myocardial protein and keratinocyte maps, from an *E. coli* 2-D gel, and from reference maps of *Spiroplasma melliterrum* and *Mycoplasma genitalium* (Sutton *et al.*, 1995; Rasmussen *et al.*, 1994; Henzel *et al.*, 1993; Cordwell *et al.*, 1995; Wasinger *et al.*, 1995), although the technique is most powerful when used in combination with another protein identification technique (Rasmussen *et al.*, 1994; Cordwell *et al.*, 1995).

MASS SPECTROMETRY SEQUENCE TAGGING

An extension of peptide mass fingerprinting has recently been described, called peptide sequence tagging (Mann and Wilm, 1994; Mann, 1995). This uses tandem mass spectrometry (MS/MS) to initially determine the mass of peptides, then subject them to fragmentation by collision with a gas, and finally determine the mass of fragments. The resulting spectra gives information about a peptide's amino acid sequence. The fragmentation masses of peptides can rarely be used to assign a complete sequence, but it usually allows a short 'sequence tag' of 2 or 3 amino acids to be determined. This sequence tag and the original peptide mass is matched by computer against a database, providing a likely identity of the peptide and the protein it came from. The major drawback for this technique as a mass screening tool is the complexity of the mass data generated and the high level of expertise required for its interpretation. Nevertheless, it represents a useful new protein identification method which greatly increases the power of peptide mass fingerprinting protein identification.

Cross-species protein identification

Protein sequence databases continue to grow at a rapid rate, yet it is not widely appreciated that close to 90% of all information contained in current protein databases comes from only 10 species (A. Bairoch, Pers. Comm.). Fortunately, this information can be used to study proteomes of organisms that are poorly defined at the molecular level, via 2-D electrophoresis and 'cross-species' protein identification (Cordwell *et al.*, 1995; Wasinger *et al.*, 1995). This approach allows proteins from reference maps of many different species to be identified without the need for the corresponding genes to be cloned and sequenced. This is particularly true for 'housekeeping' proteins, such as enzymes involved in glycolysis, DNA manipulation and protein manufacture, which are highly conserved across species boundaries. Proteins that cannot be identified across species boundaries can then become the focus of further protein characterisation and DNA sequencing efforts.

A)

Protein: APA1_HUMAN

Asx: 8.4 Cix: 19.3 Ser: 6.3 His: 1.3
 Gly: 4.2 Thr: 4.3 Ala: 6.0 Pro: 4.2
 Tyr: 2.9 Arg: 6.7 Val: 5.5 Met: 1.3
 Ile: 0.0 Leu: 15.5 Phe: 2.5 Lys: 8.8

pI Range: no range specified

Mw Range: no range specified

The closest SWISS-PROT entries are:

Rank	Score	Protein	(pI)	Mw	Description
1	0	APA1_HUMAN	5.27	28078	APOLIPOPROTEIN A-I.
2	4	APA1_MACFA	5.43	28005	APOLIPOPROTEIN A-I.
3	10	APA1_RABIT	5.15	27836	APOLIPOPROTEIN A-I.
4	14	APA1_BOVIN	5.36	27549	APOLIPOPROTEIN A-I.
5	14	APA1_CANFA	5.10	27467	APOLIPOPROTEIN A-I.
6	18	APA1_MOUSE	5.42	27922	APOLIPOPROTEIN A-I.
7	26	APA1_PIG	5.19	27598	APOLIPOPROTEIN A-I.
8	27	APA1_CHICK	5.26	27966	APOLIPOPROTEIN A-I.
9	37	DYNA_CHICK	5.44	117742	DYNACTIN, 117 KD ISOFORM.
10	39	APA4_HUMAN	5.18	43374	APOLIPOPROTEIN A-IV.

B)

Reagent: Trypsin MW filter: 10k

Scan using fragment mws of:

1953	1933	1731	1613	1401	1387
1201	1283	1252	1235	1231	1215
1031	996	873	811	813	781
732	704				

No. of database entries scanned = 72018

1	APA1_HUMAN	APOLIPOPROTEIN A-I (APO-AI). - HOMO SAPIENS
2	APA1_MACFA	APOLIPOPROTEIN A-I (APO-AI). - MACACA FASCICULARIS
3	APA1_PAPHA	APOLIPOPROTEIN A-I (APO-AI). - PAPIO HAMADRYAS
4	B41845	csf B - Treponema denticola
5	APA1_CANFA	APOLIPOPROTEIN A-I (APO-AI). - CANIS FAMILIARIS (DOG).
6	S30947	hypothetical protein 1 - Azotobacter vinelandii
7	MS2C_PEA	CHLOROPLAST HEAT SHOCK PROTEIN PRECURSOR. - PISUM SATIVU
8	S20724	Tropomyosin - African clawed frog
9	HIVV1354	HIVV1354 premature term. at 793 - Human immunodeficiency
10	TRAC_ECOLI	TRAJ PROTEIN. - ESCHERICHIA COLI.

Figure 6. Theoretical cross-species matching of human apolipoprotein A-I by amino acid composition and tryptic peptides. When an unknown protein is analysed, best ranking proteins from both techniques can be compared. If the same protein type is observed in both lists, there is high confidence in this being the identity of the unknown molecule (Cordwell *et al.*, 1995). (A) Output of ExPASy server (Appel, Bairoch and Hochstrasser, 1994) where the true amino acid composition of apolipoprotein A-I was matched against all entries in the SWISS-PROT database, without pI or MW windows. Seven of the top 10 matching proteins were apolipoprotein A-I of different species. (B) Output of MOWSE peptide mass fingerprinting program (Pappin, Hojrup and Bleasby, 1993) where true tryptic peptides of human apolipoprotein A-I were matched against the OWL database, using MW window of 10%. Four of the top ten matching proteins were apolipoprotein A-I from different species.

Rapid cross-species identification of proteins from 2-D reference maps can be undertaken with amino acid composition or peptide mass fingerprinting methods (Figure 6), but these techniques alone may not identify proteins unambiguously when phylogenetic cross-species distances are great or analysis data is of poor quality (Yates *et al.*, 1993; Shaw, 1993; Cordwell *et al.*, 1995). However, very high confidence in protein identities can be achieved when lists of best-matching proteins generated by both techniques are compared (Cordwell *et al.*, 1995; Wasinger *et al.*, 1995). The correct identification is found when the same protein is ranked highly in lists of best matches generated by both techniques. This method has allowed approximately 120 proteins from the reference map of the mollicute *Spiriplasma melliferum*, representing approximately one quarter of the proteome, to be confidently identified by reference to protein information from other species (S. Cordwell, Personal Communication). When cross-species protein identification is to be undertaken, it should be noted that the molecular weight of a protein type across species is usually highly conserved, but that protein pI can vary by more than 2 units (Cordwell *et al.*, 1995). Accurate molecular weight determination by direct mass spectrometry of proteins blotted to PVDF (Eckerskorn *et al.*, 1992) should therefore be a useful additional parameter for cross-species protein identification.

CHARACTERISATION OF POST-TRANSLATIONAL MODIFICATIONS

Many proteins are modified after translation. Such post-translational modifications, including glycosylation, phosphorylation, and sulfation (see Table 6), are usually necessary for protein function or stability. Some abnormal modifications are associated with disease (Duthel and Revol, 1993; Ghosh *et al.*, 1993; Yamashita *et al.*, 1993). In proteome studies, post-translational modifications can be examined on all proteins present, or on individual spots. Studies on all proteins provide an indication of which proteins may carry a certain type of modification. For example, 2-D gel analysis of cell cultures grown in the presence of [³H] mannose or [³²P] phosphate gives an indication of which proteins carry glycans containing mannose, and which proteins are phosphorylated (Garrels and Franza, 1989). Lectin binding studies of 2-D gels blotted to PVDF or nitrocellulose provide information on the saccharides, if any, that are carried by proteins present (Gravel *et al.*, 1994).

When individual proteins of interest carrying post-translational modifications have been found, micropreparative 2-D electrophoresis can be used to purify them in microgram quantities (Hanash *et al.*, 1991; Bjellqvist *et al.*, 1993b). If protein isoforms of similar MW and pI are to be studied, focusing with narrow range pI gradients (1 pH unit) can provide greater separation and resolution. After electrophoresis, the type and degree of protein phosphorylation can be investigated (Murthy and Iqbal, 1991; Gold *et al.*, 1994), monosaccharide composition can be determined (Weitzhandler *et al.*, 1993; Packer *et al.*, 1995), and the structure and exact site of glycoamino acids can be investigated by either Edman degradation based techniques or by mass spectrometry (Pisano *et al.*, 1993; Hubert *et al.*, 1993; Carr, Huddleston and Bean, 1993). With further development of rapid techniques, investigation of phosphorylation and monosaccharides by chromatographic or mass spectrometric means is likely to become a routine step in the characterisation of post-translational modifications of proteins from reference maps.

The
N-
com-
Ac-
mal-
each
pro-
gene
size
com-
thi-
plac-
Was
map
spe-
and

Table
PRO1
refer-
from
1996.
Speci-

Mycro-
Esche-
Surch-
Diets-
Arabi-
Caeni-
Humi-

Th
unde-
beca-
hund-
estin-
to tis-
prote-
elect-
ism
accel-
are u-
post-
diffe-
usefu-

The status of proteome projects

Many technical aspects of proteome research have already been discussed in this review, but an overview of the status of proteome projects has not yet been presented. Advances in proteome projects will initially rely on progress in genome sequencing initiatives, to enable an identity, amino acid sequence, or function to be assigned to each protein spot. Table 7 shows genome size, proteome size, and the number of proteins already defined for a number of model organisms. This indicates that whilst genome sequencing programs for *E. coli* and *S. cerevisiae* are advanced, the massive size of some other genomes (and especially the human genome) means that their complete nucleotide sequences are unlikely to be available for many years. Because of this, 2-D reference maps and proteome projects of single cell organisms like *Mycoplasma* sp., *E. coli* and *S. cerevisiae* will be the most detailed (Cordwell *et al.*, 1995; Wasinger *et al.*, 1995; Vanbogelen *et al.*, 1992; Garrels *et al.*, 1994), and complete maps of other organisms will take longer to construct. However, the use of cross-species protein identification techniques will allow proteomes of many prokaryotes and simple eukaryotes to be partially defined in reference to *E. coli* and *S. cerevisiae*.

Table 7: Estimated genome size, estimated proteome size, number of protein sequences in SWISS-PROT Release 31 (March, 1995), and approximate number of proteins of known identity on 2-D reference maps for some model organisms. Genome size data from Smith (1994), and total protein data from Bird (1995). Genome sequencing projects of *E. coli* and *S. cerevisiae* will probably be complete in 1996.

Species Name	Haploid genome size (million bp)	Estimated proteome size (total proteins)	Protein entries in SWISS-PROT	Proteins annotated on 2-D Maps
<i>Mycoplasma</i> species	0.6-0.8	400-600	100	> 100
<i>Escherichia coli</i>	4.8	4000	3170	> 300
<i>Saccharomyces cerevisiae</i>	13.5	6000	3160	> 100
<i>Dactyloscymus discoidium</i>	70	12500	202	-
<i>Arabidopsis thaliana</i>	70	14000	270	-
<i>Caenorhabditis elegans</i>	80	17800	203	-
<i>Homo sapiens</i>	2900	60000-80000	3526	> 1000

The study of vertebrate proteomes and vertebrate development is a phenomenal undertaking in comparison to the investigation of single cell organisms. This is because vast numbers of proteins are developmentally expressed, each body tissue has hundreds of unique proteins, and there are numerous tissue types. However, it is estimated that at least 35% of proteins in vertebrate cells will be conserved from tissue to tissue, constituting the 'housekeeping' proteins (Bird, 1995), with the remainder of proteins constituting a set that are specific to a cell type. Providing that standardised electrophoretic conditions are used, reference maps from many tissues of one organism can be superimposed in gel databases (e.g. Hochstrasser *et al.*, 1992). This accelerates the definition of the 'housekeeping' proteins, as well as sets of proteins that are unique to different tissue types. Such studies may, however, be complicated by post-translational modifications, which can differ on the same gene product in different tissues. Proteins that remain unknown after identification procedures will be useful in providing focus for nucleic acid sequencing initiatives.

FUTURE DIRECTIONS OF PROTEOME PROJECTS

This review has described recent advances in the area of proteome research. It has illustrated how new developments of older techniques (2-D electrophoresis and amino acid analysis) as well as the applications of new technology (mass spectrometry) have greatly widened the choice of tools the biologist and protein chemist has for the separation, identification and analysis of complex mixtures of proteins. This has made possible the establishment of detailed reference maps for organisms, which are becoming the method of choice for the definition of tissues or whole cells, and the investigation of gene expression therein.

Proteome projects are already impacting on the dogma of molecular biology that DNA sequence constitutes the definition of an organism. For example, the proteomes of different tissues of a single organism are often significantly different. Similarly, cross-species identification of proteins (for example the identification of proteins from *Candida albicans* by comparison with *S. cerevisiae*) can open up studies on organisms that are poorly molecularly defined. As cross-species identification can proceed at a pace orders of magnitude faster than a genome project in terms of defining the gene and protein complement of organisms, the need for the DNA sequencing of genomes will be avoided, and emphasis placed on those found to be novel.

Just as genome sequencing is not an end in itself, neither is an annotated 2-D protein reference map of an organism, nor indeed the identification of proteins in a proteome. So whilst an immediate aim of proteome projects is to screen proteins in reference maps, this will lead to expression studies and characterisation of post-translational modifications. The challenge that then needs to be addressed is the investigation of structure and function of proteins in a proteome. The magnitude of this is illustrated by the fact that over half the open reading frames identified in *S. cerevisiae* chromosome III were initially of no known function (Oliver *et al.*, 1992). Structural and functional studies will be an undertaking just as formidable as genome studies are now and proteome projects are becoming, but will lead to an unimaginably detailed understanding of how living organisms are constructed and how they operate.

Acknowledgements

MIRW is the recipient of an Australian Postgraduate Research Award. AAG, MIRW, IHS and KIW acknowledge assistance for proteome projects through Macquarie University Research Grants, the Australian Research Council, the Australian National Health and Medical Research Council, Beckman Instruments and GBC Scientific Equipment. DH acknowledges the financial support of a Montus Foundation Grant and the Swiss National Fund for Scientific Research (Grant # 31-33658.92). We thank colleagues who supplied work that was 'In Press' during the writing of this review.

References

- ANDERSON, N.L., HOFMANN, J.P., GEMMELL, A. AND TAYLOR, J. (1984). Global approaches to quantitative analysis of gene-expression patterns observed by use of two-dimensional gel electrophoresis. *Clinical Chemistry*, 30, 2031-2036.

- ANDERSON, N.L. AND ANDERSON, N.G. (1991). A two-dimensional gel database of human plasma proteins. *Electrophoresis*, 12, 853-906.
- ANDERSON, N.L., ESQUER-BLASCO, R., HOFMANN, J.P. AND ANDERSON, N.G. (1991). A two-dimensional gel database of rat liver proteins useful in gene regulation and drug effects studies. *Electrophoresis*, 12, 907-930.
- ANDERSON, N.L., COPPLE, D.C., BENDELE, R.A., PROBST, G.S., RICHARDSON, F.C. (1992). Covalent protein modifications and gene expression changes in rodent liver following administration of methapyrilene: a study using two-dimensional electrophoresis. *Fundamental and Applied Toxicology*, 18, 570-580.
- APPEL, R.D., BAIROCH, A. AND HOCHSTRASSER, D.F. (1994). A new generation of information retrieval tools for biologists: the example of the EXPASY WWW server. *Trends in Biochemical Sciences*, 19, 258-260.
- APPEL, R.D., HOCHSTRASSER, D.F., FUNK, M., VARGAS, J.R., PELIGRINI, C., MÜLLER, A.F. AND SCHERRER, J.R. (1991). The MELANIE project: from a biopsy to automatic protein map interpretation by computer. *Electrophoresis*, 12, 722-735.
- APPEL, R.D., SANCHEZ, J.-C., BAIROCH, A., GOLAZ, O., MIL, M., VARGAS, J.R. AND HOCHSTRASSER, D.F. (1993). SWISS-2DPAGE: a database of two-dimensional gel electrophoresis images. *Electrophoresis*, 14, 1323-1326.
- APPEL, R.D., SANCHEZ, J.-C., BAIROCH, A., GOLAZ, O., RAVIER, F., PASQUALI, C., HUGHES, G. AND HOCHSTRASSER, D.F. (1994). The SWISS-2DPAGE database of two-dimensional polyacrylamide gel electrophoresis. *Nucleic Acids Research*, 22, 3581-3582.
- BAIROCH, A. AND BOECKMANN, B. (1994). The SWISS-PROT protein sequence databank: current status. *Nucleic Acids Research*, 22, 3578-3580.
- BAKER, C.S., CORBETT, J.M., MAY, A.J., YACOB, M.H. AND DUNN, M.J. (1992). A human myocardial two-dimensional electrophoresis database: protein characterisation by microsequencing and immunoblotting. *Electrophoresis*, 13, 723-726.
- BARTLET-JONES, M., JEFFERY, W.A., HANSEN, H.F. AND PAPPIN, D.J.C. (1994). Peptide ladder sequencing by mass spectrometry using a novel, volatile degradation reagent. *Rapid Communications in Mass Spectrometry*, 8, 737-742.
- BAUER, D., MÜLLER, H., REICH, J., RIEDEL, H., AHRENKIEL, V., WARTHOF, P. AND STRAUSS, M. (1993). Identification of differentially expressed mRNA species by an improved display technique (DDRT-PCR). *Nucleic Acids Research*, 21, 4272-4280.
- BERNERS-LEE, T.J., CAILLIE, R., GROFF, J.F. AND POLLERMAN, B. (1992). *Electronic Networking: Research, Applications, and Policy*, 2, 52-58.
- BIRD, A.P. (1995). Gene number, noise reduction and biological complexity. *Trends in Genetics*, 11, 94-100.
- BJELLOVIST, B., EK, K., RIGHETTI, P.G., GIANAZZA, E., GORG, A., WESTERMEIER, R. AND POSTEL, W. (1982). Isoelectric focusing in immobilized pH gradients: principle, methodology and some applications. *Journal of Biochemical and Biophysical Methods*, 6, 317-329.
- BJELLOVIST, B., PASQUALI, C., RAVIER, F., SANCHEZ, J.-C. AND HOCHSTRASSER, D.F. (1993a). A nonlinear wide-range immobilized pH gradient for two-dimensional electrophoresis and its definition in a relevant pH scale. *Electrophoresis*, 14, 1357-1365.
- BJELLOVIST, B., SANCHEZ, J.-C., PASQUALI, C., RAVIER, F., PAQUET, N., FRITIGER, S., HUGHES, G.J. AND HOCHSTRASSER, D.F. (1993b). Micropreparative 2-D electrophoresis allowing the separation of milligram amounts of proteins. *Electrophoresis*, 14, 1375-1378.
- BJELLOVIST, B., HUGHES, G., PASQUALI, C., PAQUET, N., RAVIER, F., SANCHEZ, J.-C., FRITIGER, S. AND HOCHSTRASSER, D. (1993c). The focusing positions of polypeptides in immobilized pH gradients can be predicted from their amino acid sequences. *Electrophoresis*, 14, 1023-1031.
- BONNER, W.M. AND LASKEY, R.A. (1974). A film detection method for tritium-labeled proteins and nucleic acids in polyacrylamide gels. *European Journal of Biochemistry*, 46, 83-88.
- BOUTELL, T., GARRELS, J.I., FRANZA, B.R., MONARDO, P.J. AND LATTER, G.I. (1994). REF52 on global gel navigator: an internet-accessible two-dimensional gel electrophoresis database. *Electrophoresis*, 15, 1487-1490.
- BREWER, J., GRUND, E., HAGERLID, P., OLSSON, I. AND LIZANA, J. (1986). In *Electrophoresis '86* (M.J. Dunn, Ed.), pp. 226-229. VCH, Weinheim.

- CARR, S.A., HUDDLESTON, M.J. AND DEAN, M.F. (1993). Selective identification and differentiation of N- and O-linked oligosaccharides in glycoproteins by liquid chromatography-mass spectrometry. *Protein Science*, 2, 182-196.
- CATTRELL, K.P., MACLENNAN, D.F. AND JORGENSEN, A.O. (1983). Staining of the Ca²⁺-binding proteins, calsequestrin, calmodulin, troponin C, and S-100, with the cationic dye 'Stains-all'. *Journal of Biological Chemistry*, 258, 11267-11273.
- CELIS, J.E., CRUGER, D., KILL, J., DEJGAARD, K., LAURIDSEN, J.B., RATZ, G.P., BASSE, B., CELIS, A., RASMUSSEN, H.H., BALW, G. AND VANDEKERKHOVE, J. (1990a). A two-dimensional gel protein database of noncultured total normal human epidermal keratinocytes: identification of proteins strongly up-regulated in psoriatic epidermis. *Electrophoresis*, 11, 242-254.
- CELIS, J.E., GESSER, B., RASMUSSEN, H.H., MADSEN, P., LEFFERS, H., DEJGAARD, K., HONORE, B., OLSEN, E., RATZ, G., LAURIDSEN, J.B., BASSE, B., MOURITZEN, S., HELLERUP, M., ANDERSEN, A., WALBUM, E., CELIS, A., BALW, G., PUYE, M., VAN DAMME, J. AND VANDEKERKHOVE, J. (1990b). Comprehensive two-dimensional gel protein database offers a global approach to the analysis of human cells: the transformed amnion cells (AMA) master database and its link to genome DNA sequence data. *Electrophoresis*, 11, 898-1071.
- CELIS, J.E., RASMUSSEN, H.H., OLSEN, E., MADSEN, P., LEFFERS, H., HONORE, B., DEJGAARD, K., GROMOV, P., HOFFMANN, H.J., NIELSEN, M., VASSILEV, A., VINTERMYR, O., HAO, J., CELIS, A., BASSE, B., LAURIDSEN, J., RATZ, G.P., ANDERSEN, A.H., WALBUM, E., KJAERGAARD, I., PUYE, M., VAN DAMME, J. AND VANDEKERKHOVE, J. (1993). The human keratinocyte two-dimensional database: update 1993. *Electrophoresis*, 14, 1091-1195.
- CELIS, J.E. AND OLSEN, E. (1994). A qualitative and quantitative protein database approach identified individual and groups of functionally related proteins that are differentially regulated in simian virus 40 (SV40) transformed human keratinocytes: an overview of the functional changes associated with the transformed phenotype. *Electrophoresis*, 15, 309-344.
- CORBETT, J.M., DUNN, M.J., POSCH, A. AND GORG, A. (1994a). Positional reproducibility of protein spots in two-dimensional polyacrylamide electrophoresis using immobilized pH gradient isoelectric focusing in the first dimension - an interlaboratory comparison. *Electrophoresis*, 15, 1205-1211.
- CORBETT, J.M., WHEELER, C.H., BAKER, C.S., YACOUR, M.H. AND DUNN, M.J. (1994b). The human myocardial two-dimensional gel protein database: update 1994. *Electrophoresis*, 15, 1459-1465.
- CORDWELL, S., WILKINS, M.R., CERPA-POLJAK, A., GOOLEY, A.A., DUNCAN, M., WILLIAMS, K.L. AND HUMPHERY-SMITH, J. (1995). Cross-species identification of proteins separated by two-dimensional electrophoresis using MALDI-TOF and amino acid composition. *Electrophoresis*, 15, 438-443.
- COTTRELL, J.S. (1994). Protein identification by peptide mass fingerprinting. *Peptide Research*, 7, 115-124.
- CRIVINS, D.L., MCCOURT, D.W., THOMA, R.S., SCOTT, M.G., MACKE, K. AND SCHWARTZ, B.D. (1990). *In situ* chemical cleavage of proteins immobilized to glass-fiber and polyvinylidene fluoride membranes: cleavage at tryptophan residues with 2-(2-nitrophenylsulfenyl)-3-methyl-3-bromoindolenine to obtain internal amino acid sequence. *Analytical Biochemistry*, 187, 27-38.
- DUTHIEL, S. AND REVOL, A. (1993). Glycan microheterogeneity of alpha 1-antitrypsin in serum and meconium from normal and cystic fibrosis patients by crossed immunoelectrophoresis with different lectins (Con A, LCA, WGA). *Clinical and Chemical Acta*, 215, 173-187.
- ECKERSKORN, C., JUNGRLIT, P., MEWES, W., KLOSE, J. AND LOTTSPEICH, F. (1988). Identification of mouse brain proteins after two-dimensional electrophoresis and electroblotting by microsequence analysis and amino acid composition. *Electrophoresis*, 9, 830-838.
- ECKERSKORN, C., STRUPAT, K., KARAS, M., HILLENKAMP, F. AND LOTTSPEICH, F. (1992).

- Mass spectrometric analysis of blotted proteins after gel electrophoretic separation by matrix-assisted laser desorption/ionization. *Electrophoresis*, 13, 664-665.
- ECKERSKORN, C. AND LOTTSPEICH, F. (1993). Structural characterization of blotting membranes and the influence of membrane parameters for electroblotting and subsequent amino acid sequence analysis of proteins. *Electrophoresis*, 14, 831-836.
- ER, K., BJELLOVIST, B.J. AND RIGHETTI, P.G. (1983). Preparative isoelectric focusing in immobilized pH gradients. I. General principles and methodology. *Journal of Biochemical and Biophysical Methods*, 8, 135-155.
- FEY, S.J., CARLSEN, J., MOSE LARSEN, P., JENSEN, U.A., KJELDSEN, K. AND HALVSO, S. (1994). Two-dimensional gel electrophoresis as a tool for molecular cardiology. Proceedings of the International Society for Heart Research 'XV' European Section Meeting, pp 9-16.
- FREY, J.R., KUHN, L., KETTMAN, J.R. AND LEFKOVITS, I. (1994). The amino acid composition of 350 lymphocyte proteins. *Molecular Immunology*, 31, 1219-1231.
- GARRELS, J.I. (1989). The QUEST system for quantitative analysis of two-dimensional gels. *Journal of Biological Chemistry*, 264, 5269-5282.
- GARRELS, J.I. AND FRANZA, B.R. (1989). The REF52 protein database. *Journal of Biological Chemistry*, 264, 5283-5298.
- GARRELS, J.I., FRANZA, B.R., CHANG, C. AND LATTER, G. (1990). Quantitative exploration of the REF52 protein database: cluster analysis reveals the major protein expression profiles in responses to growth regulation, serum stimulation, and viral transformation. *Electrophoresis*, 11, 111-1130.
- GARRELS, J.I., FUTCHER, B., KOBAYASHI, R., LATTER, I., SCHWENDER, B., VOLPE, T., WARNER, J.R. AND MCLAUGHLIN, C.S. (1994). Protein identification for a *Saccharomyces cerevisiae* protein database. *Electrophoresis*, 15, 1466-1486.
- GELFI, C., BOSSI, M.L., BJELLOVIST, B. AND RIGHETTI, P.G. (1987). Isoelectric focusing in immobilized pH gradients in the pH 10-11 range. *Journal of Biochemical and Biophysical Research Methods*, 15, 41-48.
- GHARAHDAAGHI, F., ATHERTON, D., DEMOTT, M. AND MISCHE, S.M. (1992). Amino acid analysis of PVDF-bound proteins. in *Techniques in Protein Chemistry III* (R.H. Ageletti, Ed.), pp 249-260. Academic Press, San Diego.
- GHOSH, P., OKOH, C., LIU, Q.H. AND LAKSHMAN, M.R. (1993). Effects of chronic ethanol on enzymes regulating sialylation and desialylation of transferrin in rats. *Alcoholism: Clinical and Experimental Research*, 17, 576-579.
- GIOMETTI, C.S., GEMMELL, M.A., TOLLAKSEN, S.L. AND TAYLOR, J. (1991). Quantitation of human leukocyte proteins after silver staining: a study with two-dimensional electrophoresis. *Electrophoresis*, 12, 536-543.
- GIOMETTI, C.S., TAYLOR, J. AND TOLLAKSEN, S.L. (1992). Mouse liver protein database: a catalog of proteins detected by two-dimensional gel electrophoresis. *Electrophoresis*, 13, 970-991.
- GOLAZ, O., HUGHES, G.J., FRITIGER, S., PAQUET, N., BAIRICH, A., PASQUALI, C., SANCHEZ, J.C., TISSOT, J.D., APPEL, R.D., WALZER, C., BALANT, L. AND HOCHSTRASSER, D.F. (1993). Plasma and red blood cell protein maps: update 1993. *Electrophoresis*, 14, 1223-1231.
- GOLD, M.R., YUNGWIRTH, T., SUTHERLAND, C.L., INGHAM, R.J., VIANZON, D., CHIL, R., VAN OOSTVEEN, L., MORRISON, H.D. AND AEBERSOLD, R. (1994). Purification and identification of tyrosine-phosphorylated proteins from lymphocytes stimulated through the antigen receptor. *Electrophoresis*, 15, 441-453.
- GOLDBERG, H.A., DOMENICUCCI, C., PRINGLE, G.A. AND SODEK, J. (1988). Mineral-binding proteoglycans of fetal porcine calvarial bone. *Journal of Biological Chemistry*, 263, 12092-12101.
- GOOLEY, A.A., MARSHCHALEK, R. AND WILLIAMS, K.L. (1992). Size polymorphisms due to changes in the number of O-glycosylated tandem repeats in the *Dictyostelium discoideum* glycoprotein PsA. *Genetics*, 130, 749-756.
- GORG, A., POSTEL, W. AND GUNTHER, S. (1988). The current state of two-dimensional electrophoresis with immobilized pH gradients. *Electrophoresis*, 9, 531-546.
- GORG, A., POSTEL, W., GUNTHER, S., WESER, J., STRAHLER, J.R., HANASHI, S.M., SOMERLOT, L.

- AND KUICK, R. (1988). Approach to stationary two-dimensional pattern: influence of focusing time and immobilized carrier ampholyte concentrations. *Electrophoresis*, 9, 37-44.
- GRAVEL, P., GOLAZ, O., WALZER, C., HOCHSTRASSER, D.F., TURLEZ, H., AND BALANT, L.P. (1994). Analysis of glycoproteins separated by two-dimensional gel electrophoresis using lectin blotting revealed by chemiluminescence. *Analytical Biochemistry*, 221, 66-71.
- GLANTHER, S., POSTEL, W., WILRING, H. AND GORG, A. (1988). Acid phosphatase typing for breeding nematode-resistant tomatoes by isoelectric focusing with an ultranarrow immobilized pH gradient. *Electrophoresis*, 9, 616-620.
- HANASH, S.M., STRAHLER, J.R., NEEL, J.V., HAILAT, N., MELHEM, R., KEIM, D., ZHU, X.X., WAGNER, D., GAGE, D.A. AND WATSON, J.T. (1991). Highly resolving two-dimensional gels for protein sequencing. *Proceedings of the National Academy of Sciences USA*, 88, 5709-5713.
- HARRINGTON, M.G., COFFMAN, J.A., CALZONE, F.J., HOOD, L.E., BRITTEN, R.J. AND DAVIDSON, E.H. (1992). Complexity of sea urchin embryo nuclear proteins that contain basic domains. *Proceedings of the National Academy of Sciences USA*, 89, 6252-6256.
- HARRINGTON, M.G., LEE, K.H., YUN, M., ZEWERT, T., BAILEY, J.E. AND HOOD, L.E. (1993). Mechanical precision in two-dimensional electrophoresis can improve spot positional reproducibility. *Applied and Theoretical Electrophoresis*, 3, 347-353.
- HELLMAN, U., WERNSTEDT, C., GONZ, J. AND HELDIN, C.-H. (1995). Improvement of an in-gel digestion for the micropreparation of internal protein fragments for amino acid sequencing. *Analytical Biochemistry*, 224, 451-455.
- HENZEL, W.J., BILLECI, T.M., STULTS, J.T., WONG, S.C., GRIMLEY, C. AND WATANABE, C. (1993). Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. *Proceedings of the National Academy of Sciences USA*, 90, 5011-5015.
- HESS, D., COVEY, T.C., WINZ, R., BROWNSEY, R.W. AND AEBERSOLD, R. (1993). Analytical and micropreparative peptide mapping by high performance liquid chromatography/electrospray mass spectrometry of proteins purified by gel electrophoresis. *Protein Science*, 2, 1342-1351.
- HOBOMI, U., HOLTHAEVE, T. AND SANDER, C. (1994). Amino acid analysis and protein database compositional search as a rapid and inexpensive method to identify proteins. *Analytical Biochemistry*, 222, 202-209.
- HOCHSTRASSER, D.F. AND MERRIL, C.R. (1988). 'Catalysts' for polyacrylamide gel polymerization and detection of proteins by silver staining. *Applied and Theoretical Electrophoresis*, 1, 35-40.
- HOCHSTRASSER, D.F., PATCHORNIK, A. AND MERRIL, C.R. (1988). Development of polyacrylamide gels that improve the separation of proteins and their detection by silver staining. *Analytical Biochemistry*, 173, 412-423.
- HOCHSTRASSER, A.C., JAMES, R.W., PONIETTA, D. AND HOCHSTRASSER, D.F. (1991a). Preparative isoelectrofocusing and high resolution two-dimensional electrophoresis for concentration and purification of proteins. *Applied and Theoretical Electrophoresis*, 1, 333-337.
- HOCHSTRASSER, D.F., APPEL, R.D., VARGAS, R., PERKIER, R., VU RLOD, J.F., RAVIER, F., PASQUALI, C., FUNK, M., PELLIGRINI, C., MULLER, A.F. AND SCHERRER, J.R. (1991b). A clinical molecular scanner: the Melanie project. *Medical Computing*, 8, 85-91.
- HOCHSTRASSER, D.F., FRUTIGER, S., PAQUET, N., BAIROCH, A., RAVIER, F., PASQUALI, C., SANCHEZ, J.-C., TISSOT, J.-D., BJELLOVIST, B., VARGAS, R., APPEL, R.D. AND HUGHES, G.J. (1992). Human liver protein map: a reference database established by microsequencing and gel comparison. *Electrophoresis*, 13, 992-1001.
- HOLT, T.G., CHANG, C., LAURENT-WINTER, C., MURAKAMI, T., DAVIES, J.E. AND THOMPSON, C.J. (1992). Global changes in gene expression related to antibiotic synthesis in *Streptomyces hygroscopicus*. *Molecular Microbiology*, 6, 969-980.
- HONORE, B., LEFFERS, H., MADSEN, P. AND CELIS, J.E. (1993). Interferon-gamma up-regulates a unique set of proteins in human keratinocytes. Molecular cloning and expression of the cDNA encoding the RGD-sequence containing protein IGUP 1-5111. *European Journal of Biochemistry*, 218, 421-430.
- HUBERTY, M.C., VATH, J.E., YL, W. AND MARTIN, S.A. (1993). Site-specific carbohydrate

- identification in recombinant proteins using MALD-TOF MS. *Analytical Chemistry*, 65, 2791-2800.
- HUGHES, G.J., FRUTIGER, S., PAQUET, N., PASQUALI, C., SANCHEZ, J.-C., TISSOT, J.D., BAÏROCH, A., APPEL, R.D. AND HOCHSTRASSER, D.F. (1993). Human liver protein map update 1993. *Electrophoresis*, 14, 1216-1222.
- HUGHES, J.H., MACK, K. AND HAMPIRIAN, V.V. (1988). India ink staining of proteins on nylon and hydrophobic membranes. *Analytical Biochemistry*, 173, 18-25.
- JAMES, P., OLADRONI, M., CARAFOLI, E. AND GONNET, G. (1993). Protein identification by mass profile fingerprinting. *Biochemical and Biophysical Research Communications*, 195, 58-64.
- JIL, H., WHITEHEAD, R.H., REID, G.E., MORITZ, F.L., WARD, L.D. AND SIMPSON, R.J. (1994). Two-dimensional electrophoretic analysis of proteins expressed by normal and cancerous human crypts: application of mass spectrometry to peptide-mass fingerprinting. *Electrophoresis*, 15, 391-405.
- JOHNSON, R.S. AND WALSH, K.A. (1992). Sequence analysis of peptide mixtures by automated integration of Edman and mass spectrometric data. *Protein Science*, 1, 1083-1091.
- JOHNSTON, R.F., PICKETT, S.C. AND BARKER, D.L. (1990). Autoradiography using storage phosphor technology. *Electrophoresis*, 11, 355-360.
- JUNGBLUT, P., DZIONARA, M., KLOSE, J. AND WITTMANN-LEIBOLD, B. (1992). Identification of tissue proteins by amino acid analysis after purification by two-dimensional electrophoresis. *Journal of Protein Chemistry*, 11, 603-612.
- JUNGBLUT, P., OTTO, A., ZEINDL-EBERHART, E., PLEIBNER, K.-P., KNECHT, M., REGITZ-ZAGROSEK, V., FLECK, E. AND WITTMANN-LEIBOLD, B. (1994). Protein composition of the human heart: the construction of a myocardial two-dimensional electrophoresis database. *Electrophoresis*, 15, 685-707.
- KOHARA, Y., AKIYAMA, K. AND ISONO, K. (1987). The physical map of the whole *E. coli* chromosome: application of a new strategy for rapid analysis and sorting of a large genomic library. *Cell*, 50, 495-508.
- KLOSE, J. (1975). Protein mapping by combined isoelectric focusing and electrophoresis in mouse tissues. A novel approach to testing for individual point mutations in mammals. *Human Genetics*, 26, 231-243.
- LATHAM, K.E., GARRELS, J.I., CHANG, C. AND SOLTER, D. (1991). Quantitative analysis of protein synthesis in mouse embryos I: extensive re-programming at the one- and two-cell stages. *Development*, 2, 921-932.
- LATHAM, K.E., GARRELS, J.I., CHANG, C. AND SOLTER, D. (1992). Analysis of embryonic mouse development: construction of a high-resolution, two-dimensional gel protein database. *Applied and Theoretical Electrophoresis*, 2, 163-170.
- LATHAM, K.E., GARRELS, J.I. AND SOLTER, D. (1993). Two-dimensional analysis of protein synthesis. *Methods in Enzymology*, 255, 473-489.
- LE MAIRE, M., DESCHAMPS, S., MOLLER, J.V., LE CAER, J.P. AND ROSSIER, J. (1993). Electrospray ionization mass spectrometry from sodium dodecyl sulfate-polyacrylamide gel electrophoresis: application to the topology of the sarcoplasmic reticulum Ca^{2+} -ATPase. *Analytical Biochemistry*, 214, 50-57.
- LEMKIN, P.F. AND LESTER, E.P. (1989). Database and search techniques for two-dimensional gel protein data: a comparison of paradigms for exploratory data analysis and prospects for biological modelling. *Electrophoresis*, 10, 122-140.
- LEMKIN, P.F., WU, Y. AND UPTON, K. (1993). An efficient disk-based data structure for rapid searching of quantitative two-dimensional gel databases. *Electrophoresis*, 14, 1341-1350.
- LI, K.W., GERAERTS, W.P., VAN-ELK, R. AND KOOSE, J. (1989). Quantification of proteins in the subnanogram and nanogram range: comparison of the AuroDye, FerriDye, and India ink staining methods. *Analytical Biochemistry*, 182, 4-17.
- LIANG, P. AND PARDEE, A.B. (1992). Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science*, 257, 967-971.
- MIANN, M. (1995). Sequence database searching by mass spectrometric data. In *Microcharacterisation of Proteins* (R. Kellner, F. Lonspeich, and H.E. Meyer, Eds), pp 223-245. VCH, Weinheim.

- MANN, M., HOJRLUP, P. AND ROEPSTORFF, P. (1993). Use of mass spectrometric molecular weight information to identify proteins in sequence databases. *Biological Mass Spectrometry*, 22, 338-345.
- MANN, M. AND WILM, M. (1994). Error tolerant identification of peptides in sequence databases by peptide sequence tags. *Analytical Chemistry*, 66, 1390-1399.
- MATSUDAIRA, P. (1987). Sequence of picomole quantities of proteins electroblotted onto polyvinylidene difluoride membranes. *Journal of Biological Chemistry*, 262, 10035-10038.
- MONARDO, P.J., BOUTELL, T., GARRELS, J.I. AND LATTER, G.J. (1994). A distributed system for two-dimensional gel analysis. *Computer Applications in the Biosciences*, 10, 137-143.
- MORTZ, E., VORML, O., MANN, M. AND ROEPSTORFF, P. (1994). Identification of proteins in polyacrylamide gels by mass spectrometric peptide mapping combined with database search. *Biological Mass Spectrometry*, 23, 249-261.
- MURTHY, L.R. AND IOBAL, K. (1991). Measurement of picomoles of phosphoamino acids by high performance liquid chromatography. *Analytical Biochemistry*, 193, 299-303.
- MYRICK, J.E., LEMKIN, P.F., ROBINSON, M.K. AND UPTON, K.M. (1993). Comparison of the BiImage Visage 2000 and the GELLAB-II two-dimensional electrophoresis image analysis systems. *Applied and Theoretical Electrophoresis*, 3, 335-346.
- NEIDHARDT, F.C., APPLEBY, D.B., SANKAR, P., HUTTON, M.E. AND PHILLIPS, T.A. (1989). Genomically linked cellular protein databases derived from two-dimensional polyacrylamide gel electrophoresis. *Electrophoresis*, 10, 116-122.
- NIKODEM, V. AND FRESCO, J.R. (1979). Protein fingerprinting by SDS-gel electrophoresis after partial fragmentation with CNBr. *Analytical Biochemistry*, 97, 382-386.
- NOKIHARA, K., MORITA, N. AND KURIKI, T. (1992). Applications of an automated apparatus for two-dimensional electrophoresis. Model TEP-1, for microsequence analysis of proteins. *Electrophoresis*, 13, 701-707.
- O'FARRELL, P.H. (1975). High resolution two-dimensional electrophoresis of proteins. *Journal of Biological Chemistry*, 250, 4007-4021.
- O'FARRELL, P.Z., GOODMAN, H.M. AND O'FARRELL, P.H. (1977). High resolution two-dimensional electrophoresis of basic as well as acidic proteins. *Cell*, 12, 1133-1142.
- OLIVER *et al.* (1992). The complete DNA sequence of yeast chromosome III. *Nature* 357, 38-46.
- OLSEN, A.D. AND MILLER, M.J. (1988). Elsie 4: quantitative computer analysis of sets of two-dimensional gel electrophoretograms. *Analytical Biochemistry*, 169, 49-70.
- ORTIZ, M.L., CALERO, M., FERNANDEZ-PATRON, C., PATRON, C.F., CASTELLANOS, L. AND MENDEZ, E. (1992). Imidazole-SDS-Zn reverse staining of proteins in gels containing or not SDS and microsequence of individual unmodified electroblotted proteins. *FEBS Letters*, 296, 300-304.
- OSTERGREN, K., ERIKSSON, G. AND BJELLOVIST, B. (1988). The influence of support material used on band sharpness in Immobiline gels. *Journal of Biochemical and Biophysical Methods*, 16, 165-170.
- OU, K., WILKINS, M.R., YAN, J.X., GOOLEY, A.A., FUNG, Y., SHELMAK, D. AND WILLIAMS, K.L. (1995). Improved high-performance liquid chromatography of amino acids derivatised with 9-fluorenylmethyl chloroformate. *Journal of Chromatography* (in press).
- PACKER, N., WILKINS, M.R., GOLAZ, O., LAWSON, M., GOOLEY, A.A., HOCHSTRASSER, D.F., REDMOND, J. AND WILLIAMS, K.L. (1995). Characterisation of human plasma glycoproteins separated by two-dimensional gel electrophoresis. *BioTechnology* (in press).
- PAPPIN, D.J.C., HOJRLUP, P. AND BLEASBY, A.J. (1993). Rapid identification of proteins by peptide-mass fingerprinting. *Current Biology*, 3, 327-332.
- PATTERSON, S.D. (1994). From electrophoretically separated protein to identification: strategies for sequence and mass analysis. *Analytical Biochemistry*, 221, 1-15.
- PATTERSON, S.D. AND LATTER, G.J. (1993). Evaluation of storage phosphor imaging for quantitative analysis of 2-D gels using the Quest II system. *BioTechniques*, 15, 1076-1083.
- PISANO, A., REDMOND, J.W., WILLIAMS, K.L. AND GOOLEY, A.A. (1993). Glycosylation sites identified by solid-phase Edman degradation: O-linked glycosylation motifs on human glycophorin A. *Glycobiology*, 3, 429-435.
- RABILLOLD, T. (1992). A comparison between low background silver diamine and silver nitrate protein stains. *Electrophoresis*, 13, 429-439.

- RASMUSSEN, H.H., VAN DAMME, J., PIJPE, M., GESSER, B., CELIS, J.E. AND VANDEKERCKHOVE, J. (1992). Microsequences of 145 proteins recorded in the two-dimensional gel protein database of normal human epidermal keratinocytes. *Electrophoresis*, 13, 960-969.
- RASMUSSEN, H.H., MORTZ, E., MANN, M., ROEPSTORFF, P. AND CELIS, J.E. (1994). Identification of transformation sensitive proteins recorded in human two-dimensional gel protein databases by mass-spectrometric peptide mapping alone and in combination with microsequencing. *Electrophoresis*, 15, 406-416.
- RICHARDSON, F.C., HORN, D.M. AND ANDERSON, N.L. (1994). Dose-responses in rat hepatic protein modification and expression following exposure to the rat hepatocarcinogen methapyrilene. *Carcinogenesis*, 15, 325-329.
- RIGHETTI, P.G. (1990). Immobilized pH gradients: theory and methodology. In *Laboratory Techniques in Biochemistry and Molecular Biology* (R.H. Burdon and P.H. van Knippenberg, Eds) Elsevier, Amsterdam.
- RIGHETTI, P.G. AND DRYSDALE, J.W. (1973). *Annals of the New York Academy of Sciences*, 209, 163-186.
- RODRIGUEZ, L.V., GERNSTEN, D.M., RAMAGLI, L.S. AND JOHNSTON, D.A. (1993). Towards stoichiometric silver staining of proteins resolved in complex two-dimensional electrophoresis gels: real-time analysis of pattern development. *Electrophoresis*, 14, 628-637.
- ROSENFELD, J., CAPDEVIELLE, J., GUILLEMOT, J.C. AND FERRARA, P. (1992). In-gel digestion of proteins for internal sequence analysis after one- or two-dimensional gel electrophoresis. *Analytical Biochemistry*, 203, 173-179.
- SANCHEZ, J.C., RAVIER, F., PASQUALI, C., FRUTIGER, S., PAQUET, N., BJELLOVIST, B., HOCHSTRASSER, D.F. AND HUGHES, G.J. (1992). Improving the detection of proteins after transfer to polyvinylidene difluoride membranes. *Electrophoresis*, 13, 715-717.
- SANGER, F., COLLSON, A.R., HONG, G.F., HILL, D.F. AND PETERSEN, G.B. (1982). Nucleotide sequence of bacteriophage λ DNA. *Journal of Molecular Biology*, 162, 729-773.
- SCHHEEL, G.J. (1975). Two-dimensional analysis of soluble proteins. *Biochemistry*, 250, 5375-5385.
- SHAW, G. (1993). Rapid identification of proteins. *Proceedings of the National Academy of Sciences USA*, 90, 5138-5142.
- SIBBALD, P.R., SOMMERFELDT, H. AND ARGOS, P. (1991). Identification of proteins in sequence databases from amino acid composition. *Analytical Biochemistry*, 198, 330-333.
- SIMPSON, R.J., TSUGITA, A., CELIS, J.E., GARRELS, J.I. AND MEWES, H.W. (1992). Workshop on two-dimensional gel protein databases. *Electrophoresis*, 13, 1055-1061.
- SINHA, P.K., KOTTGEN, E., STOFFLER, M.-M., GIANAZZA, E. AND RIGHETTI, P.G. (1990). Two-dimensional maps in very acidic immobilized pH gradients. *Journal of Biochemical and Biophysical Methods*, 20, 345-352.
- SMITH, D.W. (1994). Introduction. In *Biocomputing: Informatics and Genome Projects* (D.W. Smith, Ed.), pp1-12. Academic Press, San Diego.
- STRUPAT, K., KARAS, M., HILLENKAMP, F., ECKERSKORN, C. AND LUTTSPEICHL, F. (1994). Matrix-assisted laser desorption ionization mass spectrometry of proteins electroblotted after polyacrylamide gel electrophoresis. *Analytical Chemistry*, 66, 464-470.
- SUTTON, C.W., PEMBERTON, K.S., COTTRELL, J.S., CORBETT, J.M., WHEELER, C.H., DUNN, M.J. AND PAPPIN, D.J. (1995). Identification of myocardial proteins from two-dimensional gels by peptide mass fingerprinting. *Electrophoresis*, 16, 308-316.
- TOLS, G.I., FAUSNAUGH, J.L., AKINYOSOYE, O., LACKLAND, H., WINTERCASH, P., VITORICA, F.J. AND STEIN, S. (1989). Amino acid analysis on polyvinylidene difluoride membranes. *Analytical Biochemistry*, 179, 50-55.
- TOVEY, E.R., FORD, S.A. AND BALDO, B.A. (1987). Protein blotting on nitrocellulose: some important aspects of the resolution and detection of antigens in complex extracts. *Journal of Biochemical and Biophysical Methods*, 14, 1-17.
- URWIN, V.E. AND JACKSON, P. (1993). Two-dimensional polyacrylamide gel electrophoresis of proteins labeled with the fluorophore monobromobimane prior to first-dimensional isoelectric focusing: imaging of the fluorescent protein spot patterns using a cooled charge-coupled device. *Analytical Biochemistry*, 209, 57-62.
- VANBOGELEN, R.A., HUTTON, M.E. AND NEIDHARDT, F.C. (1990). Gene-protein database

- of *Escherichia coli*. K-12: edition 3. *Electrophoresis*, 11, 1131-1166.
- VANBOGELEN, R.A. AND NEIDHARDT, F.C. (1991). The gene-protein database of *Escherichia coli*, edition 4. *Electrophoresis*, 12, 955-994.
- VANBOGELEN, R.A., SANKER, F., CLARK, R.L., BOGAN, J.A. AND NEIDHARDT, F.C. (1992). The gene-protein database of *Escherichia coli*, edition 5. *Electrophoresis*, 13, 101-1054.
- VANDEKERKHOVE, J., BALW, G., VANCOMPERNOLLE, K., HONORE, B. AND CELIS, J. (1990). Comparative two-dimensional gel analysis and microsequencing identifies gelsolin as one of the most prominent downregulated markers of transformed human fibroblast and epithelial cells. *Journal of Cell Biology*, 111, 95-102.
- VANFLETEREN, J.R., RAYMACKERS, J.G., VAN BUN, S.M. AND MEHUS, L.A. (1992). Peptide mapping and microsequencing of proteins separated by SDS-PAGE after limited *in situ* hydrolysis. *BioTechniques*, 12, 550-557.
- VORM, O. AND MANN, M. (1993). Improved mass accuracy in matrix-assisted laser desorption/ionization time-of-flight mass spectrometry of peptides. *Journal of the American Society for Mass Spectrometry*, 5, 955-958.
- VORM, O., ROEPSTORFF, P. AND MANN, M. (1994). Improved resolution and very high sensitivity in MALDI TOF of matrix surfaces made by fast evaporation. *Analytical Chemistry*, 66, 3281-3287.
- WALLACE, A. AND SALLZ, H.P. (1992a). Ultramicrodetection of proteins in polyacrylamide gels. *Analytical Biochemistry*, 203, 27-34.
- WALLACE, A. AND SALLZ, H.P. (1992b). Beyond silver staining. *Nature*, 357, 608-609.
- WALSH, B.J., GOOLEY, A.A., WILLIAMS, K.L. AND BRETT, S.N. (1995). Identification of macrophage activation associated proteins by two-dimensional electrophoresis and microsequencing. *Journal of Leukocyte Biology*, 57, 507-512.
- WASINGER, V.C., CORDWELL, S.J., POLJAK, A., YAN, J.X., GOOLEY, A.A., WILKINS, M.R., DUNCAN, M., HARRIS, R., WILLIAMS, K.L. AND HUMPHERY-SMITH, I. (1995). Progress with Gene-Product Mapping of the Mollicutes: *Mycoplasma genitalium*. *Electrophoresis*, 16, In Press.
- WEITZHANDLER, M., KADLECEK, D., AYDALOVIC, N., FORTE, J.G., CHOW, D. AND TOWNSEND, R. R. (1993). Monosaccharide and oligosaccharide analysis of proteins transferred to polyvinylidene fluoride membranes after sodium dodecyl sulfate-polyacrylamide gel electrophoresis. *Journal of Biological Chemistry*, 268, 5121-5130.
- WILKINS, M.R., PASQUALI, C., APPEL, R.D., OU, K., GOLAZ, O., SANCHEZ, J.-C., YAN, J.X., GOOLEY, A.A., HUGHES, G., HUMPHERY-SMITH, I., WILLIAMS, K.L. AND HOCHSTRASSER, D.F. (1995). From Proteins to Proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis. Submitted.
- WILKINS, M.R., OU, K., APPEL, R.D., GOLAZ, O., PASQUALI, C., YAN, J.X., FARNSWORTH, V., CARTIER, P., HOCHSTRASSER, D.F., WILLIAMS, K.L. AND GOOLEY, A.A. (1996). Rapid protein identification using N-terminal sequence tagging and amino acid analysis (submitted).
- WIRTH, P.J., LUO, L.D., FUJIMOTO, Y., BISGAARD, H.C. AND OLSEN, A.D. (1991). The rat liver epithelial (RLE) cell protein database. *Electrophoresis*, 12, 931-954.
- WIRTH, P.J., LUO, L.D., BENJAMIN, T., HUANG, T.N., OLSEN, A.D. AND PARMALEE, D.C. (1993). The rat liver epithelial (RLE) cell nuclear protein database. *Electrophoresis*, 14, 1199-1215.
- WU, Y., LEMKIN, P.F. AND UPTON, K. (1993). A fast spot segmentation algorithm for two-dimensional gel electrophoresis analysis. *Electrophoresis*, 14, 1351-1356.
- YAMAGUCHI, K. AND ASAKAWA, H. (1988). Preparation of colloidal gold for staining proteins electrotransferred onto nitrocellulose membranes. *Analytical Biochemistry*, 172, 104-107.
- YAMASHITA, K., IDEO, H., OHKURA, T., FUKUSHIMA, K., YUASA, I., OHNO, K. AND TAKESHITA, K. (1993). Sugar chains of serum transferrin from patients with carbohydrate deficient glycoprotein syndrome. Evidence of asparagine-N-linked oligosaccharide transfer deficiency. *Journal of Biological Chemistry*, 268, 5783-5789.
- YATES, J.R. III, SPEICHER, S., GRIFFIN, P.R. AND HUNKAPILLER, T. (1993). Peptide mass maps: a highly informative approach to protein identification. *Analytical Biochemistry*, 214, 397-408.

3
The

R.M.I.

Therm
Hamilt
Wales
Private

Introdu

Protein
dealing
type. T
structur
since th
by War
1930) i
possibl
mechar
subtilis
1986; I
comme
proteas
the char
been in
regulat
apparei
Kalisz
Barrett
Give
surpris
attracti
subject
stabilit
enzym
Sherod

* Corre
Rumoch
11261-872

Human cellular protein patterns and their link to genome DNA sequence data: usefulness of two-dimensional gel electrophoresis and microsequencing

JULIO E. CELIS,* HANNE H. RASMUSSEN,* HENRIK LEFFERS,* PEDER MADSEN,* BENT HONORÉ,* BORBALA GESSER,* KURT DEJGAARD,* JOËL VANDEKERCKHOVE*

*Institute of Medical Biochemistry and Human Genome Research Centre, Aarhus University, DK-8000 Aarhus, Denmark and *Laboratorium 100* Fysiologische Chemie, Rijksuniversiteit Gent, Belgium

ABSTRACT Analysis of cellular protein patterns by computer-aided 2-dimensional gel electrophoresis together with recent advances in protein sequence analysis have made possible the establishment of comprehensive 2-dimensional gel protein databases that may link protein and DNA information and that offer a global approach to the study of the cell. Using the integrated approach offered by 2-dimensional gel protein databases it is now possible to reveal phenotype specific protein (or proteins), to microsequence them, to search for homology with previously identified proteins, to clone the cDNAs, to assign partial protein sequence to genes for which the full DNA sequence and the chromosome location is known, and to study the regulatory properties and function of groups of proteins that are coordinately expressed in a given biological process. Human 2-dimensional gel protein databases are becoming increasingly important in view of the concerted effort to map and sequence the entire genome.—Celis, J. E.; Rasmussen, H. H.; Leffers, H.; Madsen, P.; Honoré, B.; Gesser, B.; Dejgaard, K.; Vandekerckhove, J. Human cellular protein patterns and their link to genome DNA sequence data: usefulness of two-dimensional gel electrophoresis and microsequencing. *FASEB J.* 5: 2200-2208; 1991.

Key Words: human protein patterns • 2-dimensional gel protein databases • gene expression • microsequencing • cDNA cloning • linking protein and DNA information • genome mapping and sequencing

PROTEINS SYNTHESIZED FROM information contained in the DNA orchestrate most cellular functions. The total number of proteins synthesized by a typical human cell is unknown although current estimates range from 3000 to 6000. Of these, as many as 70% may perform household functions and are expected to be shared by all cell types irrespective of their origin. There are many different cell types in the human body with perhaps 30,000 to 50,000 proteins expressed in the organism as a whole judged from the fact that about 3% of the haploid genome correspond to genes. Today only a small fraction of the total set of proteins has been identified, and little is known about the protein patterns of individual cell types or their variation under physiological and abnormal conditions.

For the past 15 years, high resolution 2-dimensional gel electrophoresis has been the technique of choice to determine the protein composition of a given cell type and for monitoring changes in gene activity through quantitative and qualitative analysis of the thousands of proteins that orchestrate various cellular functions (refs 1-6 and references

therein). The technique originally described by O'Farrell separates proteins in terms of their isoelectric point (pI) and molecular weight. Usually one chooses a condition of interest and the cell reveals the global protein behavioral response as all detected proteins can be analyzed both qualitatively and quantitatively in relation to each other. At present, most available 2-dimensional gel techniques (regular gel format) can resolve between 1000 and 2000 proteins from a given mammalian cell type, a number that corresponds to about 2 million base pairs of coded DNA. Less abundant proteins can be detected by analyzing partially purified cellular fractions.

Two-dimensional gel electrophoresis has been widely applied to analysis of cellular protein patterns from bacteria to mammalian cells (refs 1-6, and references therein). In spite of much work, however, information gathered from these studies has not reached the scientific community in its fullness because of lack of standardized gel systems and the lack of means for storing and communicating protein information. Only recently, because of the development of appropriate computer software (7-13), has it been possible to scan gels, assign numbers to individual proteins, and store the wealth of information in quantitative and qualitative comprehensive 2-dimensional gel protein databases (4, 14-23), i.e., those containing information about the various properties (physical, chemical, biological, biochemical, physiological, genetic, immunological, architectural, etc.) of all the proteins that can be detected in a given cell type. Such integrated 2-dimensional gel protein databases offer an easy and standardized medium in which to store and communicate protein information and provide a unique framework in which to focus a multidisciplinary approach to study the cell. Once a protein is identified in the database, all of the information accumulated can be easily retrieved and made available to the researcher. In the long run, protein databases are expected to foster a wide variety of biological information that may be instrumental to researchers working in many areas of biology—among others, cancer and oncogene studies, differentiation, development, drug development and testing, genetic variation, and diagnosis of genetic and clinical diseases (Fig. 1).

The approach using systematic 2-dimensional gel protein analysis has recently gained a new dimension with the advent of techniques to microsequence major proteins recorded

*To whom correspondence should be addressed, at: Institute of Medical Biochemistry and Human Genome Research Centre, Ole Worms Alle, Bldg. 170, University Park, DK-8000 Aarhus C, Denmark.

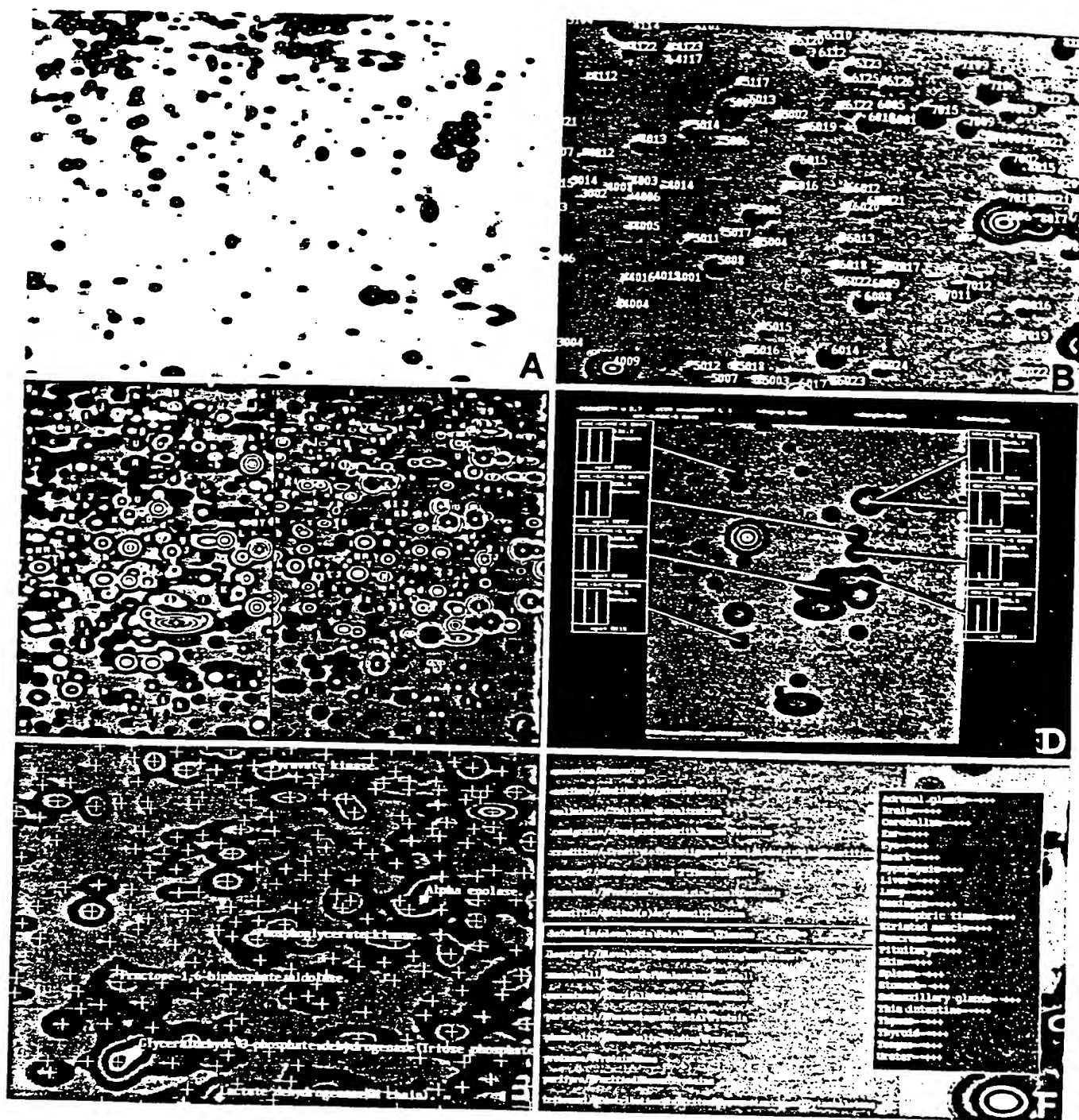


Figure 2. *A*) Synthetic image of a fraction of an IEF gel of the master image of AMA cellular proteins. *B*) As in *A* but showing numbers assigned to each spot. *C*) Comparison of AMA (left) and normal human embryonal lung MRC-5 fibroblasts (right) IEF proteins patterns. Matched proteins are indicated by a + or by the same letters in both gels. Once a protein is matched, information contained in the various categories available in the master AMA database can be transferred. *D*) Synthetic image of a fraction of an IEF fluorogram of [³⁵S]methionine labeled proteins from normal human MRC-5 fibroblasts. The histograms show levels of synthesis of a few proteins in MRC-5 (left bar) and SV40 transformed MRC-5 (right bar) fibroblasts. *E*) Polypeptides that contain information under the category glycolytic pathway. *F*) The function peruse annotation for spot allows the operator to inquire about categories and information available for a given protein. *G*) Relative abundance of cytoskeletal and cytoskeletal-related proteins in quiescent, proliferating, and SV40-transformed MRC-5 fibroblasts. *H*) Polypeptides that contain information under the category partial amino acid sequences.

cross-matched experiments (18, 22).

Once a standard map of a given protein sample is made, one can enter qualitative annotations to make a reference database. Our master 2-dimensional gel database of transformed human amnion cell (AMA) proteins (20) lists 3430 polypeptides of which 2592 correspond to cellular components, having pI's ranging from 4 to 13 and molecular weights between 8.5 and 230 kDa. The most abundant proteins in the database correspond to total actin (3.87% of total protein; about 90 million molecules per cell) while the lesser abundant of the recorded polypeptides are present in the vicinity of 5000 molecules per cell. Some annotation categories we are using to establish the master AMA database include: 1) protein identification (comigration with purified proteins, 2-dimensional immunoblotting, microsequencing); 2) amounts (total amounts and levels of synthesis); 3) subcellular localization (nuclear, cytoskeletal, membrane, membrane receptors, specific organelles, etc.); 4) antibodies; 5) posttranslational modifications (phosphorylation, glycosylation, methylation etc.); 6) microsequencing; 7) cell cycle specificity (specific variations in levels of synthesis and amount); 8) regulatory behavior (effect of hormones, growth factors, heat shock, etc.); 9) rate of synthesis in normal and transformed cells (proliferation sensitive proteins, cell cycle specific proteins, oncogenes, components of the pathway (or pathways) that control cell proliferation); 10) function (mainly from comigration with proteins of known function); 11) sets of proteins that are coordinately regulated (hierarchy of controls, differential gene expression in various cells, etc.); 12) cDNAs (cloned cDNAs); 13) proteins that are specific to a given disease (systematic comparison of protein patterns of fibroblast proteins from healthy and diseased individuals); 14) expression and exploitation of transfected cDNAs; 15) pathways (metabolic, others); 16) gene localization (genetic and physical); 17) effect of microinjected antibody on patterns of protein synthesis; and 18) secreted proteins.

Information entered for any spot in a given annotation category can be easily retrieved by asking the computer to display the information on the color screen. For example, Fig. 2E shows a synthetic image of a NEPHGE gel (master AMA database) displaying the information contained under the entry glycolytic pathway. Alternatively, one can use the function peruse annotations for spot to directly ask the computer to list all the entries available for a particular protein. By clicking the mouse in a given entry (in this case, presence in fetal human tissues) it is possible to take a quick look at the information in that particular entry (Fig. 2F).

A major obstacle encountered in building comprehensive 2-dimensional gel protein databases is identifying the large number of proteins separated by this technology. In our databases (20, 21), known proteins are identified by one or a combination of the following procedures: 1) comigration with known proteins, 2) 2-dimensional gel immunoblotting using specific antibodies, and 3) microsequencing of Coomassie Brilliant Blue stained human proteins recovered from dried 2-dimensional gels (see next section). Protein identification by means of microsequencing may be difficult, as individual protein members of families with short peptide differences may escape detection. In the gene-protein database of *E. coli* K-12 (14, 23), another major 2-dimensional gel database available at present, proteins are being identified by a wider range of tests that include comigration with purified proteins; genetic criterion (deletion, insertion, frameshift, nonsense, missense, regulatory), plasmid-bearing strains and in vitro synthesis of protein; selective labeling (methylation, phosphorylation); peptide map similarity; and physiological criterion and selective derivatization.

So far we have received nearly 550 antibodies from laboratories all over the world and these are being systematically tested by 2-dimensional gel immunoblotting for antigen determination. Similarly, purified proteins and organelles provided by several laboratories have greatly aided identification of unknown proteins (20, 21). We routinely request antibodies and protein samples and promise the donors to make available all the information we may have accumulated on that particular protein. For example, Table 1 lists entries available for Lipocortin V (IEF SSP 8216), also known as annexin V, VAC- α endonexin II, renocortin, chromobindin-5', anticoagulant protein, PAP-I, γ -calcimedlin, IBC, calphobindin, and anchorin CII.

As mentioned previously, one distinct advantage of 2-dimensional gel electrophoresis is the possibility of studying quantitative variations in cellular protein patterns that may lead to identification of groups of proteins that are expressed coordinately during a given biological process. Quantitation, however, is not an easy task as reflected by the lack of published data on global cellular protein patterns. We believe this is partly due to difficulties in obtaining sets of gels that are suitable for computer analysis (streaking, material remaining at the origin, etc.) as well as to limitations (laborious editing time, need of calibration strips to merge images, limited dynamic range, etc.) in the computer analysis systems available at the moment. Perhaps the most advanced quantitative studies published so far using computer analysis have been carried out by Garrels and co-workers (18, 22). In particular, these investigators have established a quantitative rat protein database (18, 22) designed to study growth control (proliferation, growth inhibitors, and stimulation) and transformation in well-defined groups of cell lines obtained by transformation of rat REF52 cells with SV40, adenovirus, and the Kirsten murine sarcoma virus. These studies have revealed clusters of proteins induced or repressed during growth to confluence as well as groups of transformation-sensitive proteins that respond in a differential fashion to transformation by DNA and RNA viruses. A most interesting feature of this quantitative database is the discovery of a group of coregulated proteins that show similar expression patterns as the cell cycle-regulated DNA replication protein known as proliferating cell nuclear antigen (PCNA)/cyclin (45).

In our human databases, most quantitations have been carried out by estimating the radioactivity contained in the polypeptides by direct counting of the gel pieces in a scintillation counter (20, 21). Up to 700 proteins can be cut out through appropriate exposed films in a period of time comparable to that required for editing a synthetic image. Manual quantitation of this large number of spots is difficult without the assistance of a master reference image and a numbering system that can be used to identify the spots. Using this approach, we have recorded quantitative changes in the relative abundance of 592 [35 S]methionine-labeled proteins synthesized by quiescent, proliferating, and SV40 transformed human embryonic lung MRC-5 fibroblasts (21). Some data concerning cytoskeletal and cytoskeletal-related proteins are presented in Fig. 2G. Our studies as well as those of Garrels and co-workers (18, 22) may in the long run help define patterns of gene expression that are characteristic of the transformed state.

OTHER 2-DIMENSIONAL GEL PROTEIN DATABASES

As mentioned previously there are other 2-dimensional gel databases available in computer form that have been pub-

phocytes, leukocytes, leukemic cells) mouse (NIH/3T3 cells, T lymphocytes), *Aplysia*, yeast (*Saccharomyces cerevisiae*), plants (wheat, barley, sorghum), and *Euglena*. Databases of tissue protein, (brain, whole mouse, liver) and body fluid proteins (plasma proteins, cerebrospinal fluid, urine, and milk) are being established in several laboratories. The reader is directed to the review by Celis et al. (4) for details and references concerning these databases.

MICROSEQUENCING HAS ADDED A NEW DIMENSION TO COMPREHENSIVE 2-DIMENSIONAL GEL DATABASES: A DIRECT LINK BETWEEN PROTEINS AND GENES

The development of highly sensitive amino acid gas-phase or liquid-phase sequencers (24), together with the establishment of efficient protein and peptide sample preparation methods, has opened the possibility to perform a systematic sequence analysis of proteins resolved by 2-dimensional gel electrophoresis. Indeed, generated pieces of protein sequences can be used to search for protein identity (comparison with available sequences stored in databanks) as well as for preparing specific DNA probes for cloning of as yet uncharacterized proteins (Fig. 1). In addition, partial protein sequences can be stored in 2-dimensional gel databases (for example, see Fig. 2H) and offer a unique link between proteins and genes (Fig. 1).

In the early 1970s gel electrophoresis was used to purify proteins for sequencing purposes (reviewed by Weber and Osborn in ref 25). Proteins were recovered by diffusion and sequenced by the manual dansyl-Edman degradation at the nanomole level. This technique was further refined by using electro-elution to recover proteins and by miniaturizing the system (26). This method has been used extensively, but showed increasing drawbacks (low yields, protein samples contaminated by free amino acids, and NH_2 -terminal blocking) as the amounts of handled protein gradually became smaller (e.g., at the 10 picomol level).

Most of the problems referred to above have been minimized with the introduction of protein-electroblotting procedures (27-32). When proteins are blotted on chemically inert membranes, it is possible to sequence the immobilized proteins directly without additional manipulations. Thus, depending on the amount of bound protein and its nature, this direct sequencing procedure generally yields NH_2 -terminal sequences containing 10-40 residues. As such, this technique was used to identify, by their NH_2 -terminal sequences, differentially expressed major proteins from total cellular extracts separated on 2-dimensional gels. A major difficulty encountered in this procedure is the occurrence of frequent artefactual blockage of the proteins. Several studies suggest that this phenomenon is mainly due to reaction with contaminants (particularly unpolymerized acrylamide present in the gel) and to a high dilution of the protein (low concentration of the protein per unit membrane surface). In addition to this primarily technical problem, many proteins are blocked in vivo by acylation or by a pyrrolidone carboxylic acid cap.

The problem of partial or complete NH_2 -terminal blockage can be circumvented by generating internal amino acid sequences. This is achieved by fragmenting the protein present in the gel (gel in situ cleavage) or by cleaving it while bound to the membrane (membrane in situ cleavage) (33-35). In both cases, proteins are either cleaved in a restricted way (e.g., by limited enzymatic digestion or by using restriction chemical cleavage conditions) or fragmented into smaller peptides.

Of the different combinations examined, we had good results by using exhaustive proteolytic digestion on membrane-immobilized proteins. This method has been described for Ponceau red-stained proteins on nitrocellulose blots (34), for Amido-black-stained Immobilon-bound proteins, and for fluorescamine-detected proteins on glass fiber membranes (35). The proteases used (trypsin, chymotrypsin, or pepsin) cleave at multiple sites, generating small peptides that elute from the blot into the digestion buffer from which they are purified by reversed-phase high performance liquid chromatography (HPLC) before being sequenced individually. Although each of these manipulations could be expected to result in a reduced yield of final sequence information, we were surprised that the peptides could be sequenced with high efficiency. In our hands, this approach could be routinely applied to gel-purified proteins available in amount ranging from 5 to 10 μg , and often yielded sequence information covering more than 30% of the total protein. As membrane-immobilized proteins are not homogeneously digested, but rather show protease sensitivity next to resistant regions, the number of peptides generated is much lower than expected from the number of potential cleavage sites. Consequently, HPLC peptide chromatograms are less complex and most peptides can be recovered in pure form.

As only limited amounts of a protein mixture can be loaded on a 2-dimensional gel, proteins of interest are often obtained in yields insufficient for the currently available sequencing technology. More material can be obtained by enriching for a certain subcellular fraction (purified cell organelles) or by exploiting affinity (dyes, metals, drugs, etc) or hydrophobic properties of proteins before gel analysis. All of the sequencing results accumulated so far in the human protein database (20) (a few are shown in Fig. 2H) have been obtained from analysis of protein spots collected from 2-dimensional gels that had been stained with Coomassie blue according to standard procedures and dried for storage. Proteins are recovered from the collected gel pieces by a protein-elution-concentration device, combined with gel electrophoresis and electroblotting. Details of this technique have been reported in a previous communication (42) and a brief outline is given below.

Combined gel pieces are allowed to swell in gel sample buffer (a total volume of 1.5 ml). The gel pieces combined with the supernatant are then collected into a large slot made in a new gel. The slot is further filled with Sephadex G-10 equilibrated in gel sample buffer. During consecutive gel electrophoresis, most of the electrical current passes on the side of the slot instead of passing through the slot. This results in both a vertical stacking and horizontal contraction of the protein band. With this device the protein is efficiently eluted from the gel pieces and concentrated from a large volume into a narrow spot. The highly concentrated (about 5 mm^2) protein spot is then electroblotted on PVDF-membranes, stained with Amido black, and in situ digested with trypsin. The peptides generated during digestion elute from the membrane into the supernatant, and can be separated by narrow bore reversed-phase HPLC and collected individually for sequence analysis.

Using this and previous procedures (37, 39, 42), we have so far analyzed 70 protein spots collected from 2-dimensional gels (20, and unpublished observations) (see for example Fig. 2H). The sequence information amounts to 2100 allocated residues corresponding to an average of 30 residues per protein spot. So far we have made cDNAs of many of the unknown proteins that have been microsequenced, and a substantial number has been cloned and sequenced. All available information indicates that it may be possible to obtain partial sequence information from most of

21. Celis, J. E., Dejgaard, K., Madsen, P., Leffers, H., Gesser, B., Honoré, B., Rasmussen, H. H., Olsen, E., Lauridsen, J. B., Ratz, G., Mouritzen, S., Hellerup, M., Andersen, A., Walbum, E., Celis, A., Bauw, G., Puype, M., Van Damme, J., and Vandekerckhove, J. (1990) The MRC-5 human embryonal lung fibroblast two-dimensional gel cellular protein database: quantitative identification of polypeptides whose relative abundance differs between quiescent, proliferating and SV40 transformed cells. *Electrophoresis* 12, 1072-1113
22. Garrels, J. I., Franza, B. R., Chang, C., and Latter, G. (1990) Quantitative exploration of the REF52 protein database: cluster analysis reveals the major protein expression profiles in responses to growth regulation, serum stimulation, and viral transformation. *Electrophoresis* 12, 1114-1130
23. Van Bogelen, R. A., Hutton, M. E., and Neidhardt, F. C. (1990) Gene-protein database of *Escherichia coli* K-12. 3rd ed. *Electrophoresis* 12, 1131-1166
24. Hewick, R. M., Hunkapiller, M. W., Hood, L. E., and Dreyer, W. J. (1981) A gas-liquid solid phase peptide and protein sequencer. *J. Biol. Chem.* 256, 7990-7997
25. Weber, K., and Osborn, M. (1985) In *The Proteins and Sodium Dodecyl Sulfate: Molecular Weight Determination on Polyacrylamide Gels and Related Procedures* (Neurath, H. et al., eds) Vol. 1, pp. 179-223. Academic, New York
26. Hunkapiller, M. W., Lujan, E., Ostrander, F., and Hood, L. E. (1983) Isolation of microgram quantities of proteins from polyacrylamide gels for amino acid sequence analysis. *Methods Enzymol.* 91, 227-236
27. Vandekerckhove, J., Bauw, G., Puype, M., Van Damme, J., and Van Montagu, M. (1985) Protein-blotting on polybrene-coated glass-fiber sheets. *Eur. J. Biochem.* 152, 9-19
28. Aebersold, R. H., Teplov, D. B., Hood, L. E., and Kent, S. B. H. (1986) Electroblooming onto activated glass. *J. Biol. Chem.* 261, 4229-4238
29. Bauw, G., De Loose, M., Inzé, D., Van Montagu, M., and Vandekerckhove, J. (1987) Alterations in the phenotype of plant cells studied by NH₂-terminal amino acid-sequence analysis of proteins electroblotted from two-dimensional gel-separated total extracts. *Proc. Natl. Acad. Sci. USA* 84, 4806-4810
30. Matsudaira, P. (1987) Sequence from picomole quantities of proteins electroblotted onto polyvinylidene difluoride membranes. *J. Biol. Chem.* 262, 10035-10038
31. Eckerskorn, C., Mewes, W., Goretzki, H., and Lottspeich, F. (1985) A new siliconized-glass fiber as support for protein-chemical analysis of electroblotted proteins. *Eur. J. Biochem.* 176, 509-519
32. Moos, M., Jr., Nguyen, N. Y., and Liu, T.-Y. (1988) Reproducible high yield sequencing of proteins electrophoretically separated and transferred to an inert support. *J. Biol. Chem.* 263, 6003-6008
33. Kennedy, T. E., Gawinowicz, M. A., Barzilai, A., Kandel, E. R., and Sweatt, J. D. (1988) Sequencing of proteins from two-dimensional gels by using in situ digestion and transfer of peptides to polyvinylidene difluoride membranes: application to protein associated with sensitization in *Aplysia*. *Proc. Natl. Acad. Sci. USA* 85, 7008-7012
34. Aebersold, R. H., Leavitt, J., Saavedra, R. A., Hood, L. E., and Kent, S. B. H. (1987) Internal amino acid sequence analysis of protein separated by one- or two-dimensional gel electrophoresis after in situ protease digestion on nitrocellulose. *Proc. Natl. Acad. Sci. USA* 84, 6970-6973
35. Bauw, G., Van Den Bulcke, M., Van Damme, J., Puype, M., Van Montagu, M., and Vandekerckhove, J. (1988) Protein electroblotting on polybase-coated glassfiber and polyvinylidene difluoride membranes: an evaluation. *J. Prot. Chem.* 7, 194-199
36. Celis, J. E., Ratz, G. P., Madsen, P., Gesser, B., Lauridsen, J. B., Leffers, H., Rasmussen, H. H., Nielsen, H. V., Cruger, D., Basse, B., Honoré, B., Møller, O., Celis, A., Vandekerckhove, J., Bauw, G., Van Damme, J., Puype, M., and Van Den Bulcke, M. (1989) Comprehensive, human cellular protein databases and their implication for the study of genome organization and function. *FEBS Lett.* 244, 247-254
37. Bauw, G., Van Damme, J., Puype, M., Vandekerckhove, J., Gesser, B., Lauridsen, J. B., Ratz, G. P., and Celis, J. E. (1989) Protein-electroblotting and -microsequencing strategies in generating protein databases from two-dimensional gels. *Proc. Natl. Acad. Sci. USA* 86, 7701-7705
38. Aebersold, R., and Leavitt, J. (1990) Sequence analysis of proteins separated by polyacrylamide gel electrophoresis. Towards an integrated protein database. *Electrophoresis* 11, 517-527
39. Bauw, G., Rasmussen, H. H., Van Den Bulcke, M., Van Damme, J., Puype, M., Gesser, B., Celis, J. E., and Vandekerckhove, J. (1990) Two-dimensional gel electrophoresis, protein electroblotting and microsequencing: a direct link between proteins and genes. *Electrophoresis* 11, 528-536
40. Tempst, P., Link, A. J., Riviere, L. R., Fleming, M., and Ellicone, C. (1990) Internal sequence analysis of protein separated on polyacrylamide gels at the submicrogram level: improved methods, applications and gene cloning strategies. *Electrophoresis* 11, 537-553
41. Eckerskorn, C., and Lottspeich, F. (1990) Combination of two-dimensional gel electrophoresis with microsequence and amino acid composition analysis: improvement of speed and sensitivity in protein characterization. *Electrophoresis* 11, 554-561
42. Rasmussen, H. H., Van Damme, J., Bauw, G., Puype, M., Gesser, B., Celis, J. E., and Vandekerckhove, J. (1991) In *Methods in Protein Sequence Analysis* (Jörnvall, H., and Höög, J. O., eds) pp. 103-114. Eighth International Conference on Methods in Protein Sequence Analysis. Birkhäuser Verlag, Boston
43. Olson, A. D., and Miller, M. J. (1988) Elsie 4: quantitative computer analysis of sets of two-dimensional gel electrophoretograms. *Anal. Biochem.* 169, 49-70
44. Vincens, P., Paris, N., Pujol, J. L., Gaboriaud, C., Rabilloud, T., Penneret, J., Matherat, P., and Tarroux, P. (1986) HERMeS: a second generation approach to the automatic analysis of two-dimensional electrophoresis gels. Part I: Data acquisition. *Electrophoresis* 7, 347-356
45. Celis, J. E., Madsen, P., Celis, A., Nielsen, H. V., and Gesser, B. (1987) Cychn (PCNA, auxiliary protein of DNA polymerase- δ) is a central component of the pathway(s) leading to DNA replication and cell division. *FEBS Lett.* 220, 1-7
46. Anderson, N. G., and Anderson, N. L. (1982) The human protein index. *Clin. Chem.* 28, 739-748

Bo Franzén¹
Stig Linder²
Ken Okuzawa³
Harabumi Kato³
Gert Auer¹

¹Division of Tumor Pathology,
Department of Pathology, Division
of Experimental Oncology,
Karolinska Hospital and Institute,
Stockholm Sweden
²Tokyo Medical College, Department
of Surgery, Tokyo
³Division of Experimental Oncology,
Karolinska Hospital and Institute,
Stockholm

Nonenzymatic extraction of cells from clinical tumor material for analysis of gene expression by two-dimensional polyacrylamide gel electrophoresis

We have compared different methods of preparation of malignant cells for two-dimensional electrophoresis (2-DE). We found all methods using fresh tissue to be superior compared to methods using frozen tissue. Our results indicate that nonenzymatic methods of preparation of tumor cells, including fine needle aspiration, scraping and squeezing, have advantages over methods using enzymatic extraction of cells. Nonenzymatic methods are rapid, appear to reduce loss of high molecular protein species, and alleviate the necessity of separating viable and nonviable cells by Percoll gradient centrifugation. Using these techniques, high-quality 2-DE maps were derived from tumors of the lung and breast. In the resulting polypeptide patterns, heat shock proteins, non-muscle tropomyosins and intermediate filament were identified. We conclude that nonenzymatic extraction of malignant cells from fresh tumor tissue improves the possibilities that these techniques may be useful in clinical diagnosis.

1 Introduction

Tumors may develop by a number of different mechanisms in any given cell type. At the time of diagnosis, tumors will have progressed along different pathways to various stages of malignancy. To provide a basis for individual therapy it is of importance to examine specific properties of the tumor cell population in each patient. A large number of different markers have been described in order to increase the diagnostic accuracy. It is likely that a combination of several markers is needed in the future in order to reflect different properties of the tumor. One important method for the resolution of a large number of potential markers is two-dimensional electrophoresis (2-DE). Extensive efforts are being made in identifying various polypeptides separated by 2-DE and to characterize how the expression of these polypeptides is affected by the response to cellular transformation and various culture conditions [1,2]. It would be of value to transfer this information to 2-DE separations of polypeptides from tumor tissue samples. However, one prerequisite is that the quality of the 2-DE gels from tumor samples is comparable in quality with 2-DE gels from samples of cultured cells.

Frozen tumor tissues are commonly used for various biochemical assessments. However, if such samples are analyzed by 2-D polyacrylamide gel electrophoresis (PAGE), the polypeptide patterns are obscured by contamination of serum- and connective tissue proteins. Such nontumor-cell-related variations represent serious problems in the interpretation and inter-patient comparison of 2-DE

patterns [3]. 2-DE patterns of cells prepared from fresh tumor material were analyzed after enzymatic extraction of tumor cells [4, 5] or after culturing tumor fragments in medium containing radioactive amino acids [6]. These procedures may, however, lead to alterations in the gene expression/polypeptide patterns. We are only aware of one study where nonenzymatic extraction of cells from fresh tumor tissue (prostate cancer) was used to prepare samples for 2-D PAGE [4]. We have examined enzymatic extraction and various nonenzymatic preparation techniques, including fine needle aspiration, for the preparation of cells from fresh tumor tissues. We describe nonenzymatic extraction procedures that are rapid, lead to high-quality 2-DE patterns, and that alleviate the necessity to purify tumor cell populations from dead cells.

2 Materials and methods

2.1 Cell cultures and samples used for spot identification

A rat embryonal fibroblast cell line, WT2 (a kind gift from Dr. J. I. Garrels and Dr. S. Patterson) was used for the identification of a number of heat shock and structural proteins. Human normal diploid lung fibroblasts, WI38, human epithelial breast carcinoma cells, MDA-231 and MCF-7 were purchased from ATCC and grown as recommended. Polypeptides prepared from a leukemia type pre-B-ALL were separated by 2-DE. The 2-DE map was then analyzed by Dr. S. M. Hanash (University of Michigan, Ann Arbor, USA).

2.2 Tumor tissues samples

In this study, 2-DE maps from seven tumors were used as representative illustrations: two adenocarcinoma of the lung (LA, and LB, mucinous, both cases intermediate grade of differentiation), one squamous carcinoma of the lung (LS), one carcinoid-like breast cancer (BC), one microfollicular adenoma (highly differentiated) of the thyroid (TA), one highly differentiated hyperneph-

Correspondence: Dr. Bo Franzén, Division of Tumor Pathology, Department of Pathology, L1:01, Karolinska Hospital and Institute, 10401 Stockholm 60, Sweden

Abbreviations: 2-DE, Two-dimensional polyacrylamide gel electrophoresis; IEF, isoelectric focusing; LDH, lactate dehydrogenase; NP-40, Nonidet P-40; PBS, phosphate buffered saline; PCNA, proliferating cell nuclear antigen; PIH, protease inhibitors; PMSF, phenylmethyl sulfonyl fluoride; SDS, sodium dodecyl sulfate; WW, wet weight

PMSF (0.2 mM), EDTA (1.0 mM), 0.5% Nonidet P-40 (NP-40), and 3-[3-cholamido propyl]-dimethylammonio]-1-propane sulfonate (CHAPS; 25 mM) was added carefully, mixed for 2.5 h and centrifuged for 15 min at

10000 rpm to remove any insoluble material. Duplicate or triplicate samples were taken for protein determination [11]. Samples were stored at -80°C prior to isoelectric focusing (IEF).

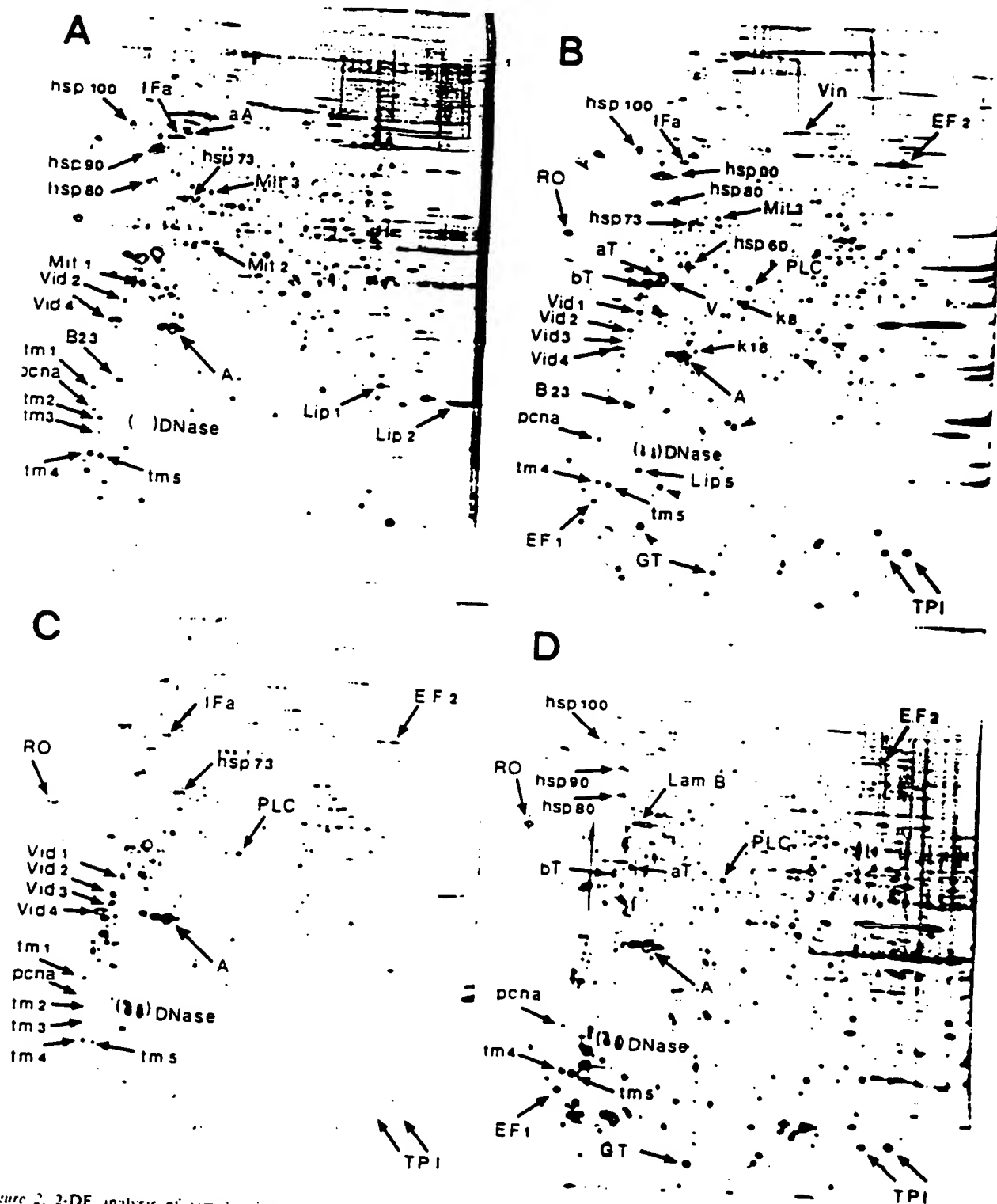


Figure 2. 2-DE analysis of samples from three cell lines and one leukemia used for the identification of polypeptides: (A) WT2; (B) MDA-231, arrowheads mark some low molecular weight cytosolic polypeptides; (C) W138 and (D) pre B-ALL. The abbreviations for identified spots are explained in Table 1.

cells (Fig. 2d). Polypeptides were identified through a laboratory exchange of cell samples/2-DE maps and through 2-DE analysis of purified proteins (Table 1).

3.2 Preparation of samples from solid tumors

3.2.1 Fresh versus frozen tissue

An adenocarcinoma of the lung (LA) was prepared for 2-DE by conventional methods using frozen material (Fig. 3a). There are several possibilities for the poor resolution using frozen tissue, including the presence of high molecular weight protein aggregates. Filtering extracts through 0.1 μ m filters (Durapore, Millipore) resulted in a slightly improved resolution (not shown). When fresh tumor tissue from tumor LA was used for sample preparation, using fine needle aspiration to collect the cells, the resolution was considerably improved (Fig. 3b). The use of fresh tissue resulted in a general increase in resolution, which was most pronounced in the 50–100 kDa molecular mass range. A number of differences in the protein profiles of the gels in Figs. 3a and 3b can be observed, some of which are indicated in the figures. The decrease in serum albumin in Fig. 3b is likely to result from loss of serum proteins occurring when cells were pelleted after aspiration. Other differences, such as the decreased level of transformation-sensitive tropomyosins (TM1-TM3), may result from enrichment of tumor cells in the sample of Fig. 3b. Fine needle aspiration, a well-established technique in cytology, extracts mainly tumor cells because of decreased intercellular adhesiveness of neoplastic cells as compared to normal tissue. Microscopic examination of Diff-Quick-stained extracted cells from case LA revealed almost 100% tumor cells, whereas the whole tissue extract contained approximately 60% tumor cells.

Table 1. Names and abbreviations for identified spots

Spot	Name	Basis for identification
A	Actins	a
aA	α -Actinin	a
B23	Protein B23 /Numatrin	a
EF2	Elongation factor 2	a
EF1	Elongation factor 1 B	a
GT	Glutathione-S-transferase (<i>p</i>)	a
hsp60	Heat shock protein 60	a
hsp73	Heat shock protein 73	a
hsp80	Heat shock protein 80, GRP78, BIP	a
hsp90	Heat shock protein 90	a
hsp100	Heat shock protein 100, Endoplasmic	a
IFa	Intermediate filament associated	a
k8	Cytokeratin 8	b and a
Lamb	Lamin B	a
Lip1	Lipocortin I	a
Lip2	Lipocortin II	a
Lip5	Lipocortin V	a
Mit1	Mitcon 1/8 – F1 ATPase	a
Mit2	Mitcon 2	a
Mit3	Mitcon 3	a
MRP	Mucine Related Polypeptides	—
pcna	Proliferating cell nuclear antigen	c and a
PLC	Phospholipase C (1)	a
RO	RO/SS-A antigen	a
SA	Serum Albumin	b and a
aT	α -Tubulin	a
bT	β -Tubulin	a
tm1	Non-muscle tropomyosin isoform 1	b and a
tm2	Non-muscle tropomyosin isoform 2	b and a
tm3	Non-muscle tropomyosin isoform 3	b and a
tm4	Non-muscle tropomyosin isoform 4	b and a
tm5	Non-muscle tropomyosin isoform 5	b and a
TP1	Triose phosphate isomerase	a
V	Vimentin	b and a
Vid1	Vimentin derived protein	b and a
Vid2	Vimentin derived protein	b and a
Vid3	Vimentin derived protein	b and a
Vid4	Vimentin derived protein	b and a
Vin	Vinculin	a

a. homologous position with respect to other mammalian systems
b. purified protein(s)
c. immunoblotting

A

B

Figure 4. 2-DE analysis of a case of breast carcinoma (BC). Comparison of 2-DE quality and some differences in detected spots (arrow heads indicate increased intensity and circles or bracket indicate decreased intensity of the same spots) between (A) enzymatically and (B) nonenzymatically (scraped) tissue preparation.

difference in intensity were lower than when a nonenzymatic preparation was compared with an enzymatic preparation.

2-DE maps of satisfactory quality were prepared by a third procedure. Cells were released from small pieces of tumor by squeezing (see Section 2). Some examples of this are shown in Fig. 6 where 2-DE maps derived from a case of hypernephroma, KH (Fig. 6a), a case of thyroid tumor, TA (Fig. 6b) and a case of corpus cancer, CP (Fig. 6c) can be seen. We conclude that nonenzymatic techniques are useful for 2-DE analysis of a number of different tumors. The quality of the resulting gels is com-

parable to that obtained using cultured cells (compare the gels in Fig. 2 with those in Fig. 4, 6 and 7). Which of these methods will be optimal will, in our experience, depend on the tumor material. For example, very small tumors are preferably extracted by squeezing; on the other hand, breast cancers (which are often fibrous) yield satisfactory samples using scraping.

3.2.3 Purification of cells on percoll gradients

We considered the possible advantage of separating viable cells from dead cells, erythrocytes, and debris using discontinuous Percoll gradients. Cells collected

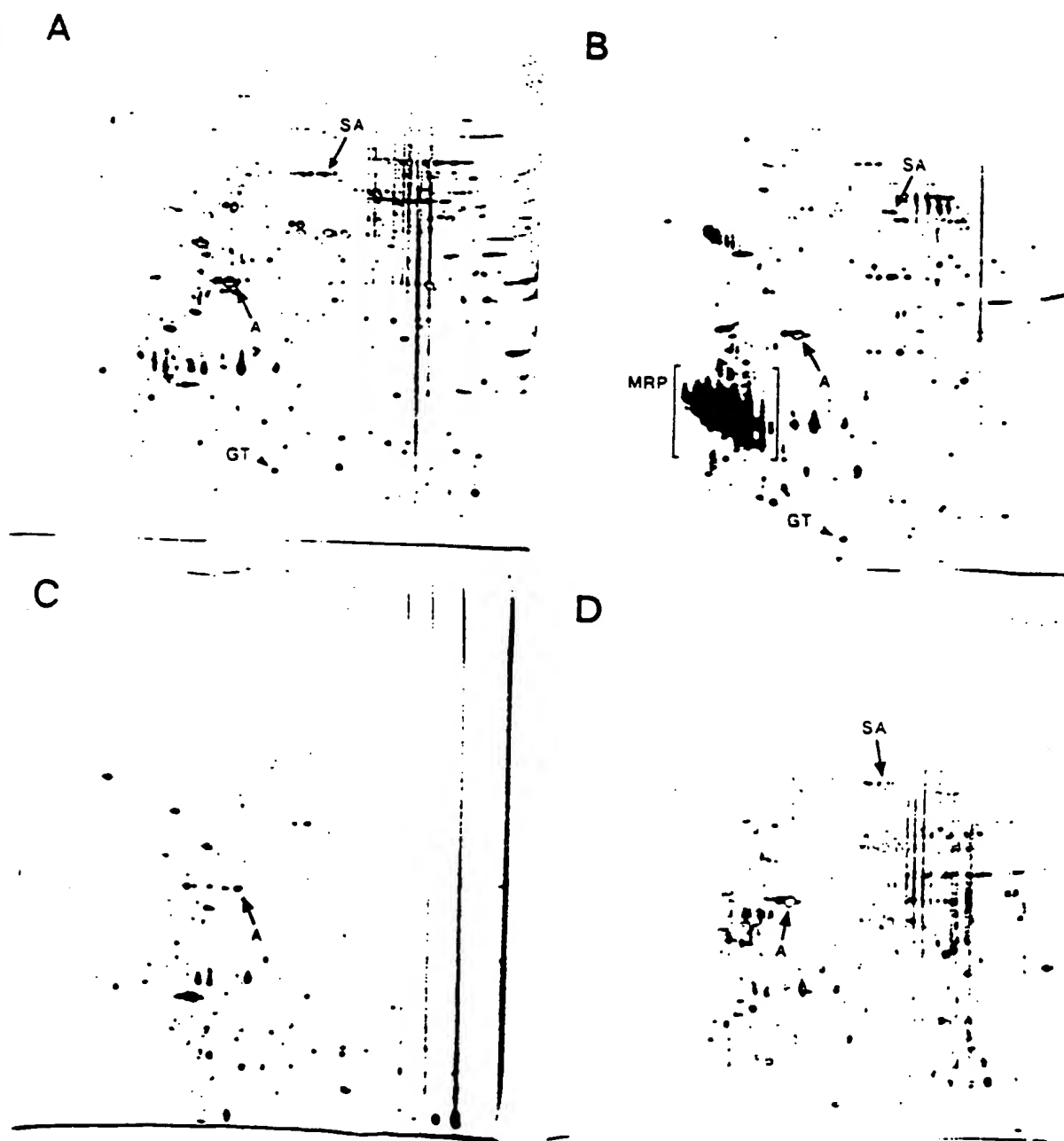


Figure 7. 2-DE analysis of polypeptides from viable (b and d) and nonviable (a and c) cells of an adenocarcinoma of the lung (LB), separated using discontinuous Percoll density gradient. Nonenzymatic preparation technique (a and b) and enzymatic preparation technique (c and d) are compared.

with these observations (Fig. 8). A number of potential and interesting markers, like tropomyosin isoforms, cytokeratins and heat shock proteins, appear to be insensitive to loss of viability during the preparation procedure. We have to date made numerous observations of alterations in the expression of these polypeptides in breast cancers and lung cancers.

Another problem that may occur, irrespective of sample preparation techniques used, is admixture of lymphocytes. These cases are easily detectable in smears and it may therefore be possible to select lymphocyte specific spots as "internal markers" for the 2-D PAGE analysis. Studies using this approach are in progress. Many of the polypeptides identified are structural (Table 1). Since the expression of many of these polypeptides are known to vary between normal and malignant cells, the possibility to determine their expression simultaneously is appealing. In the specific case of breast cancer, alterations in the expression of intermediate filament proteins (cytokeratins) are known to occur during tumor progression [23]. Other proteins known to be differentially expressed between normal cells and transformed cells are tropomyosins, numatrin/B23, heat shock proteins and PCNA. To this end, we have observed alterations in the expression of cytokeratin 8, hsp 90, and non-muscle tropomyosin isoform 2 during malignant progression. (Okuzawa *et al.*, in preparation and Franzen *et al.*, in preparation).

The method of choice for sample preparation from tumor tissues will depend on the properties of the tumor material studied. It may be important to use only one method when comparing cases within one group, as differences were observed between methods. The advantages of the nonenzymatic techniques are (i) that it minimizes contamination with connective tissue, (ii) that problems with contamination of serum proteins are avoided, and (iii) that separation of viable and dead cells is not necessary. Hereby the resolving power of 2-D PAGE is maximized for the analysis of human tumors and studies on inter-tumor variations in gene expression are facilitated. In addition, the polypeptide patterns obtained may be more representative for the *in vivo* tumor cell since the use of enzymes and incubations have been minimized.

We would like to thank Dr. J. I. Garrels, Dr. S. Pattersson, Dr. S. M. Hanash and Dr. J. E. Celis for making sample and 2-DE map exchanges possible. This study was sup-

ported by grants from the Swedish Cancer Society and the Cancer Society in Stockholm.

Received March 5, 1993

5 References

- [1] Celis, J. E., Dejgaard, K., Madsen, P., Leffers, H., Gesser, B., Honore, B., Rasmussen, H. H., Olsen, E., Lauridsen, J. B. and Ratz, G., *Electrophoresis* 1990, 11, 1072-1113.
- [2] Garrels, J. I., Franza, B. R., Chang, C., Latter, G., *Electrophoresis* 1990, 11, 1114-1130.
- [3] Franzen, B., Iwabuchi, H., Kato, H., Lindholm, J. and Auer, G., *Electrophoresis* 1991, 12, 509-515.
- [4] Sherwood, E. R., Berg, L. A., Mitchell, N. J., McNeal, J. E., Kozlowski, J. M. and Lee, C., *J. Urology* 1990, 143, 167-171.
- [5] Endler, A. T., Young, D. S., Wold, L. E., Lieber, M. M. and Curie, R. M., *J. Clin. Chem. Clin. Biochem.* 1986, 24, 981-992.
- [6] Forchhammer, J. and Macdonald-Bravo, H., in: Celis, J. E. and Bravo, R., (eds.), *Gene Expression in Normal and Transformed Cells*, Plenum, New York 1983, pp. 291-314.
- [7] Linder, S., Brzeski, H. and Ringertz, N. R., *Exp. Cell. Res.* 1979, 120, 1-14.
- [8] Celis, J. E. and Bravo, R. (Eds.), *Two-dimensional Gel Electrophoresis of Proteins*, Academic Press, New York 1984, pp. 3-36.
- [9] Garrels, J. I., *J. Biol. Chem.* 1979, 254, 7961-7977.
- [10] Anderson, N. L., *Two-Dimensional Electrophoresis, Operation of the ISO-DALT System*, Large Scale Biology Press, Washington, DC 1988, 162.
- [11] Bradford, M., *Anal. Biochem.* 1976, 72, 248.
- [12] Tracy, R. P., Wold, L. E., Currie, L. M. and Young, D. S., *Clin. Chem.* 1982, 28, 890-899.
- [13] Merrill, C. R., Goldman, D., Sedman, S. A. and Elbert, H. M., *Science* 1981, 211, 1437-1438.
- [14] Morrissey, J. H., *Anal. Biochem.* 1981, 117, 307-310.
- [15] Gard, D. L., Bell, P. B., Lazarides, E., *Proc. Natl. Acad. Sci. USA* 1979, 76, 3894-3898.
- [16] Matsumura, F., Lin, J.-C., Yamashiko-Matsumura, S., Thomas, G. P. and Topp, W. C., *J. Biol. Chem.* 1983, 258, 13954-13960.
- [17] Paulin, D., Forest, N. and Perreau, J., *J. Mol. Biol.* 1980, 144, 95-101.
- [18] Blobel, G. A., Moll, R., Franke, W. W., Kayser, K. W. and Gould, V. E., *Am. J. Pathol.* 1985, 121, 235-247.
- [19] Ochs, D. C., McConkey, H. E. and Guard, N. L., *Exp. Cell. Res.* 1981, 135, 355-362.
- [20] Bhattacharya, B., Guddamanuga, L. P., Valverius, E. M., Salomon, D. S. and H. L. Cooper, *Cancer Res.* 1990, 50, 2105-2112.
- [21] Sommers, C. L., Walker-Jones, D., Heckford, S. E., Worland, P., Valverius, A., Clark, R., McCornick, F., Stampfer, M., Abularch, S. and Gelmann, E. P., *Cancer Res.* 1989, 49, 425A-4263.
- [22] Trask, D. K., Bond, V., Zajchowski, D. A., Yaswen, P., Suh, T. and Sager, R., *Proc. Natl. Acad. Sci. USA* 1990, 87, 2319-2323.
- [23] Trask, D. K., Bond, V., Zajchowski, D. A., Yaswen, P., Suh, T. and Sager, R., *Proc. Natl. Acad. Sci. USA* 1990, 87, 2319-2323.

Bengt Bjellqvist*
Bodil Basse
Eyðfinnur Olsen
Julio E. Celis

Institute of Medical Biochemistry
and Danish Centre for Human
Genome Research, Aarhus
University, Aarhus

Reference points for comparisons of two-dimensional maps of proteins from different human cell types defined in a pH scale where isoelectric points correlate with polypeptide compositions

A highly reproducible, commercial and nonlinear, wide-range immobilized pH gradient (IPG) was used to generate two-dimensional (2-D) gel maps of [³⁵S]methionine-labeled proteins from noncultured, unfractionated normal human epidermal keratinocytes. Forty one proteins, common to most human cell types and recorded in the human keratinocyte 2-D gel protein database were identified in the 2-D gel maps and their isoelectric points (pI) were determined using narrow-range IPGs. The latter established a pH scale that allowed comparisons between 2-D gel maps generated either with other IPGs in the first dimension or with different human protein samples. Of the 41 proteins identified, a subset of 18 was defined as suitable to evaluate the correlation between calculated and experimental pI values for polypeptides with known composition. The variance calculated for the discrepancies between calculated and experimental pI values for these proteins was 0.001 pH units. Comparison of the values by the *t*-test for dependent samples (paired test) gave a *p*-level of 0.49, indicating that there is no significant difference between the calculated and experimental pI values. The precision of the calculated values depended on the buffer capacity of the proteins, and on average, it improved with increased buffer capacity. As shown here, the widely available information on protein sequences cannot, *a priori*, be assumed to be sufficient for calculating pI values because post-translational modifications, in particular N-terminal blockage, pose a major problem. Of the 36 proteins analyzed in this study, 18–20 were found to be N-terminally blocked and of these only 6 were indicated as such in databases. The probability of N-terminal blockage depended on the nature of the N-terminal group. Twenty six of the proteins had either M, S or A as N-terminal amino acids and of these 17–19 were blocked. Only 1 in 10 proteins containing other N-terminal groups were blocked.

1 Introduction

As compared with carrier ampholyte isoelectric focusing (CA-IEF), the application of immobilized pH gradients (IPGs) in the first dimension in 2-D gel electrophoresis offers improved reproducibility [1] because the nature of the pH gradient makes the resulting focusing positions insensitive to the focusing time [2] and to the type of sample applied [3]. The recently introduced ready-made IPG strips [4] seem to be an ideal substitute for the carrier ampholyte gradients, which until now have been the most commonly used first dimensions in 2-D gel electrophoresis. The availability of standardized first dimensions opens the possibility of comparing 2-D gel maps of various cell types generated in different laboratories, provided that the focusing positions of a number of easily recognizable polypeptide spots common to the cell types

in question are known. Even though this approach is limited to experiments performed with the same standardized IPG, the flexibility provided by IPGs allows the pH gradient to be adjusted to the requirements of a particular experiment.

Exchange and communication of 2-D gel protein data requires a pH scale that is independent of the particular IPG used and by which the results can be described. The introduction of carbamylation trains and the relation of focusing positions to the spots in these trains represented a step forward towards solving the reproducibility problem experienced with carrier ampholyte focusing [5]. Problems associated with the use of carbamylation trains were mainly due to lack of temperature control and to the use of nonequilibrium focusing conditions. Accordingly, the pattern variation involved not only the resulting pH gradients, but also the relative spot positions as related to each other and to spots in the carbamylation trains. Even though the question of reproducibility has, to a large extent, been solved, the carbamylation trains are still not ideal as markers because the spots in the trains do not represent defined entities but rather a large number of differently carbamylated peptides having close pI values. As a result, the spots are large and poorly defined as compared to the ordinary polypeptide spots in 2-D gel maps.

Correspondence: Professor J. E. Celis, Institute of Medical Biochemistry and Danish Centre for Human Genome Research, Aarhus University, DK-8000 Aarhus C, Denmark

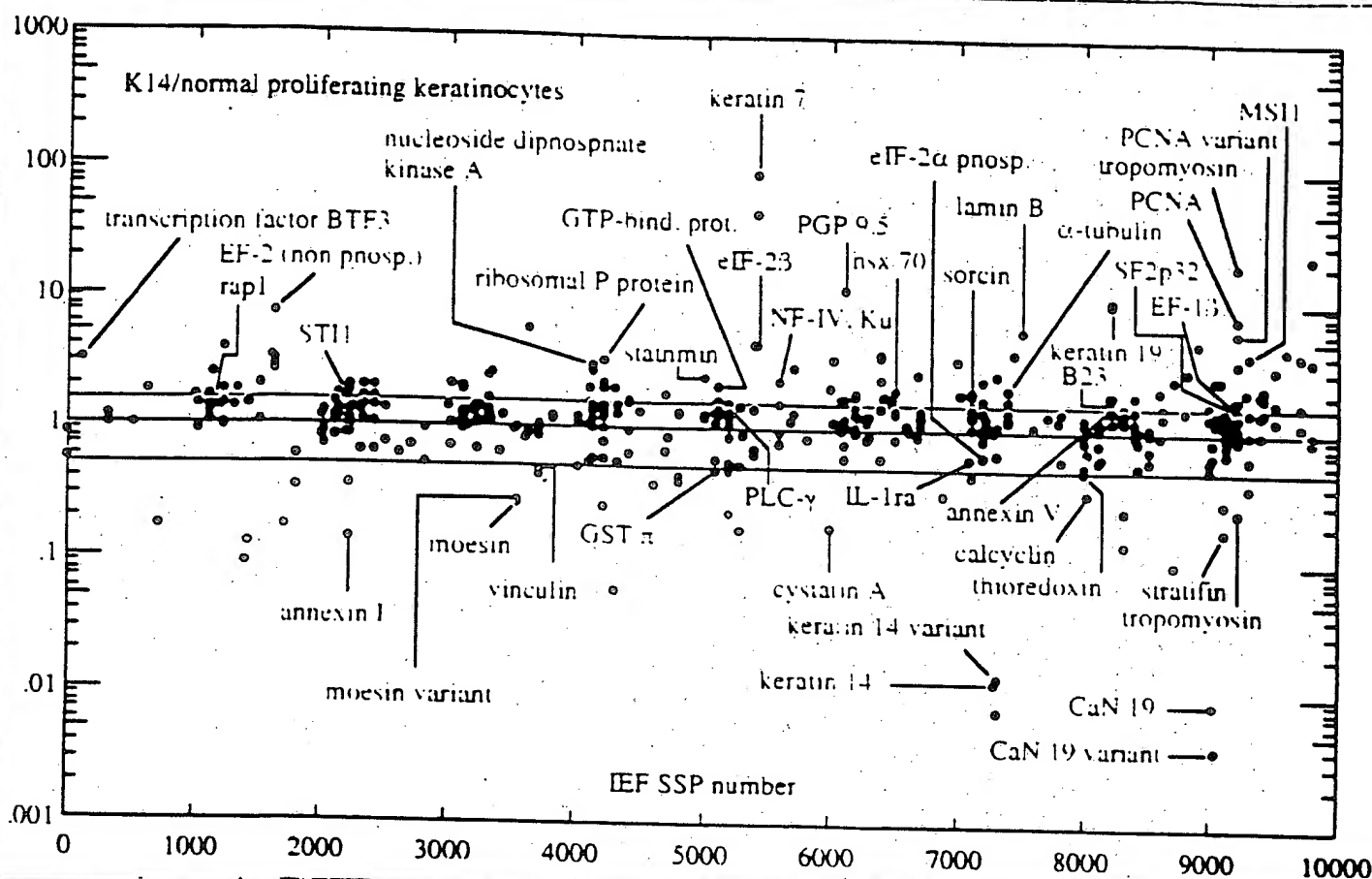
Abbreviations: CA-IEF, carrier ampholyte-isoelectric focusing; SSP, sample spot number

* Present address: Pharmacia Biotech AB, S-751 82 Uppsala, Sweden

ELECTROPHORESIS

An International Journal

3-4'94



PAPER SYMPOSIUM

ELECTROPHORESIS IN CANCER RESEARCH

Guest Editor: Julio E. Celis



Neidhardt *et al.* [6] defined the pH gradient in 2-D gel experiments by pI markers whose pI values were calculated from the amino acid composition. Focusing positions of other polypeptides could be predicted from their composition but the pK values needed for the pI calculations were unknown. Various groups employing this approach do not use the same pK values [6, 7] and therefore, the pI values derived in this way cannot be expected to describe the variation of the hydrogen ion activity. In spite of this fact, it is still possible to make approximate predictions of focusing positions because the pK values used to define the pH gradient are also used to calculate pI values and to predict the focusing positions. Errors in pK assignments are therefore compensated. A pH scale which correctly reflects the variation in hydrogen ion activity during focusing should improve the precision of the predictions, but this has never been implemented with CA-IEF focusing as a first dimension in 2-D gel electrophoresis. The main reason for this are the problems associated with pH measurements in focused gels containing high concentrations of urea.

IPGs can be described from the concentration variation of the immobilized groups, provided that the pK values of these groups are known for the conditions prevailing during focusing. To avoid measurements on gels, Gianazza *et al.* [8] suggested the use of pK values derived by addition of determined pK shifts. Recently, direct determinations of pK differences between immobilized groups in IPGs were made by determining pI-pK values in overlapping narrow-range IPGs [9, 10] and the results verified the applicability of the Gianazza approach. A description of the focusing results in a pH scale, which correctly describes the variation of the hydrogen ion activity for the focusing conditions used, not only allows the comparison of 2-D gel maps generated with different IPGs, but also opens the possibility for correlating the focusing position of a polypeptide with its composition [9]. Experiments by Bjellqvist *et al.* [9, 10] have implied that pH scales showing good correlation between calculated and experimental pI values can be derived for any of the conditions commonly used for focusing in connection with 2-D gel electrophoresis. These pH scales are then defined through the pK values of the immobilized groups in the IPG containing gel. To be useful for interlaboratory comparisons, however, the pH scale has to be defined through pI values of easily recognizable spots present in the 2-D gel map. So far, pI determinations in a useful pH scale, combined with determinations of pK values needed for pI calculations, have only been made for the pH range 4.5–6.5 at 10°C [9]. CA-IEF focusing as described by O'Farrell [11] does not control the temperature of the first dimension, which can be expected to be slightly above room temperature. With IPGs, the temperature commonly used is about 20°C [4, 12] or 25°C [13] and this is a critical parameter that needs to be controlled [14].

The present work was designed to compare 2-D gel maps of different cell types in a laboratory applying both CA-IEF and IPG focusing at a common temperature. To this end we have generated 2-D gel maps of proteins from noncultured, unfractionated normal human epidermal keratinocytes with IPG in the first dimension

and a focusing temperature of 25°C. We have used commercial nonlinear, wide-range IPG strips which give 2-D gel maps that are closely similar to the ones resulting with the CA-IEF technique used to establish the human keratinocyte database [15]. As an initial step towards interlaboratory comparisons of results obtained with the nonlinear gradient as a first dimension we report here on the focusing positions of 41 known proteins that are common to most human cell types. The pH range covered corresponds to the range in classical CA-IEF 2-D gel electrophoresis and in order to use these proteins as internal standards for comparing 2-D gel maps generated with other IPGs we determined their pI values with narrow-range IPGs in the first dimension. We have compared the calculated *versus* experimental pI values and show that it is necessary to have further information (absence or presence and nature of posttranslational modifications), in addition to amino acid composition to be able to calculate pI values that correspond to the actual experimental values. The pK values used for the calculations are provided and the usefulness of pI prediction in relation to database information is discussed. Furthermore, we comment on the possibility of using experimentally determined pI values to verify the available database information on polypeptide composition.

2 Materials and methods

2.1 Apparatus and chemicals

Equipment for isoelectric focusing and horizontal SDS electrophoresis (Multiphor[®] II electrophoresis chamber, Immobiline[®] strip tray, Multidrive XL programmable power supply, Macrodrive power supply and Multitemp[®] II) was from Pharmacia LKB Biotechnology AB (Uppsala, Sweden). Vertical second-dimensional gels were run in the home-made equipment described in [15]. The IPG strips with the wide-range nonlinear pH gradient were either Immobiline DryStrip[®] pH 3–10 NL, 180 mm or alternatively 160 mm long IPG strips with a corresponding pH gradient. In both cases the IPG strips were delivered by Pharmacia LKB. Immobiline, Pharmalyte, Ampholine, GelBond as well as PAG film and the ready-made horizontal SDS gels (ExcelGel[®] XL SDS 12–14) were also from Pharmacia LKB. Purified proteins and peptides were from Sigma (St. Louis, MO).

2.2 Sample preparation

Preparation and labeling of unfractionated keratinocytes as well as fibroblasts have been described in [16]. Cells were lysed in a solution containing 9.8 M urea, 2% w/v NP-40, 100 mM DTT and 2% v/v Ampholine pH 7–9.

2.3 2-D gel electrophoresis

First-dimensional focusing was performed according to Görg *et al.* [2] with some minor modifications, as described in [9]. Rehydration of the IPG strips was made in a solution containing 9.8 M urea, 2% w/v CHAPS, 10 mM DTT and 2% v/v carrier ampholyte mixture. The carrier ampholyte mixture consisted of 2 parts Pharmalyte

4–6.5. 1 part Ampholine pH 6–8 and 1 part Pharmalyte pH 8–10.5. Usually, cathodic sample application was used and the samples were diluted 2–20 times in a solution containing 9.8 M urea, 4% w/v CHAPS, 1% w/v DTT and 35 mM Tris base. For acidic application, the Tris-base was substituted with 100 mM acetic acid. The degree of dilution and sample volume (20–100 μ L) depended on the particular sample and the IPG, and whether visualization of the proteins was to be done by Coomassie Brilliant Blue or silver staining. With the wide-range non-linear IPG, 10–30 μ g of total protein was loaded for silver staining and 100–200 μ g for Coomassie staining. Focusing was done overnight with Vh products in the range of 45–60 kVh with 160 mm long strips and 50–70 kVh with 180 mm long strips. Solubilization of polypeptides and blocking of -SH groups prior to the second-dimensional run, as well as loading on the second-dimensional gel was done as described in [9]. The stacking gel was omitted and 5–10 mm were left at the top of the second-dimensional gel for applying the IPG strip. The space was filled with electrode buffer containing 0.5% w/v agarose. Casting, running, staining and autoradiography were carried out as described in [15].

2.4 Experimental determination of pI values

The determination of the pK differences between Immobilines pK 4.6, pK 6.2 and pK 7.0 necessary for the calibration of the pH scale at 25°C in 9.8 M urea was done as described in [9] with the same narrow-range IPGs. The pH scale was defined by setting the pK value of Immobiline pK 4.6 equal to 4.61 [9] and the determined pK differences gave the pK values of Immobilines pK 6.2 and pK 7.0, equal to 5.73 and 6.54, respectively. The pK differences found are in good agreement with values derived from [17] and [8] by extrapolation to 9.8 M urea concentration. As in [9], additional narrow-range recipes have been used for determining pI values. With narrow-range IPGs extending to pH values higher than the pK value of Immobiline pK 7.0, anodic sample application was used with acetic acid added to the sample solution. Otherwise, cathodic sample application was used with the same sample buffer as for wide-range IPGs.

2.5 Protein compositions used for pI calculations

With the exception of vimentin, protein compositions are from the Swiss-Prot database [18]. For vimentin, we used the data from [19], where the amino acid at position 41 is a D instead of a S. Information in the Swiss-Prot database on phosphorylation has been disregarded because it was known from earlier studies (J. E. Celis, unpublished results) that the spots in question corresponded to the unphosphorylated forms of the peptides.

2.6 Calculation of pI values

For the pI calculations it was assumed that the same pK value could be used for an amino acid residue in all polypeptides and in all positions in the peptide except for *N*- or *C*-terminally placed amino acids. For the pK values of the *N*-terminal amino groups the effect of the

different substituents on the α -carbon were taken into account. The calculations of pI values were made with the aid of the IPG-maker program [20].

2.7 pK values used for pI calculations

For the carboxyl terminal group and internal glutamyl and aspartyl residues the same pK values were used as in [9]. For *C*-terminal glutamyl and aspartyl residues, separate pK values were derived with the aid of the Taft equations [9, 21]. The pK values of histidyl groups were calculated from the pI values of human carbonic anhydrase I as in [9]. For *N*-terminal glycine a pK value of 7.50 was used. The pK shift caused by a substituent on the α -carbon was assumed to be identical with the pK shift the substituent caused for the amino group in the amino acid, i.e. 2.28 pH units were subtracted from the pK values for the amino groups in the amino acids given in [22, 23]. The approximate pK value of 9 for the cystenyl group was taken from [24]. For tyrosyl and arginyl groups we used the pK values for the amino acids [22, 23]. For lysyl groups the effect of high urea concentration on amino groups was taken into account and 0.5 pH units were subtracted from the amino acid pK value. These last three pK values are far from the pH range under study and the results found would have been the same if lysyl and arginyl groups were assumed to be fully ionized while the ionization of tyrosyl groups were neglected. A complete list of the pK values used is given in Table 1.

Table 1. pK Values used for the ionizable groups in peptides
9.8 M urea, 25°C

Ionizable group	pK
<i>C</i> -terminal	3.55
<i>N</i> -terminal	
Ala	7.50
Met	7.00
Ser	6.93
Pro	8.36
Thr	6.82
Val	7.44
Glu	7.70
Internal	
Asp	4.05
Glu	4.45
His	5.98
Cys	9
Tyr	10
Lys	10
Arg	12
<i>C</i> -terminal side chain groups	
Asp	4.55
Glu	4.75

2.8 Statistical analysis

Statistical comparisons of the experimental and calculated pI values were done on an Apple Macintosh IIsi using the statistical package Statistica/Mac, release 3.0b (from StatSoft Inc., Tulsa, Oklahoma). Calculated and experimental pI values were compared by the *t*-test for

correlated samples (paired *t*-test). The normality of *pI* differences was estimated graphically by probability plots. The variances of the data presented here and the similar data on plasma and liver proteins in [9] were compared by the *F*-test.

3 Results and discussion

3.1 Identification of polypeptides and *pI* determinations

The 2-D gel maps of [³⁵S]methionine-labeled proteins from noncultured, unfractionated normal human kerati-

nocytes, focused with the nonlinear, wide-range IPG and CA-IEF pH gradients in the first dimension, are shown in Figs. 1 and 2, respectively. The IPG extends to higher *pH* values but otherwise the two patterns are very similar and most of the spots in the IPG pattern can be directly related to the corresponding spots in the CA-IEF gel. To obtain comparable patterns it was important to keep the focusing temperature as similar as possible. Compared to other studies [1-4, 9, 10, 12-14], we increased the urea concentration in the focusing gel to 9.8 M because keratins streaked badly in the focusing dimension when 8 M urea was used, presumably due to

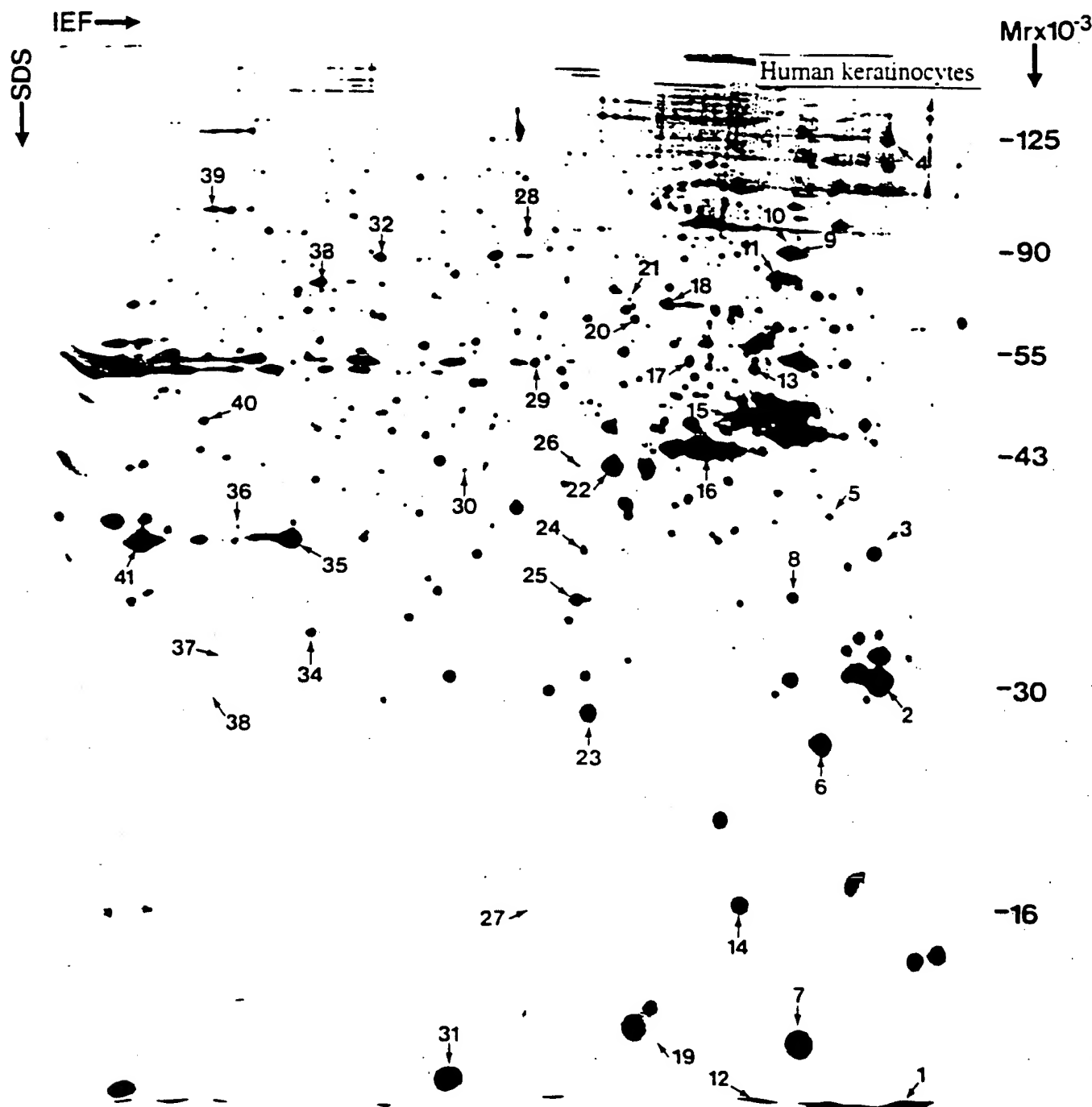


Figure 1. 2-D gel protein map of [³⁵S]methionine-labeled proteins from noncultured, unfractionated normal human keratinocytes focused with the nonlinear, wide-range IPG in the first dimension. The position of the 41 proteins analyzed in this study is indicated.

aggregates of acidic and basic keratins. An increase in urea concentration to 9 M or more eliminated these streaks; apart from this effect, no other major changes in the focusing positions were observed. In Fig. 1 we have indicated the positions of 41 known proteins from the human keratinocyte 2-D gel database that are most likely common to most human cell types. The choice was made because these proteins are easy to identify with certainty. With the exception of stratifin (spot 2), involucrin (spot 4) and keratin 14 (spot 15), which are all

epithelial markers, these proteins are also present in human fibroblasts (Fig. 3) and lymphocytes (results not shown), and therefore can be used as landmarks for comparing 2-D gel maps derived from different cell types. In Table 2 the 41 proteins are listed together with their sample spot numbers (SSP) in the human keratinocyte protein database and pI values determined in 2-D gel maps generated with narrow-range IPGs in the first dimension.

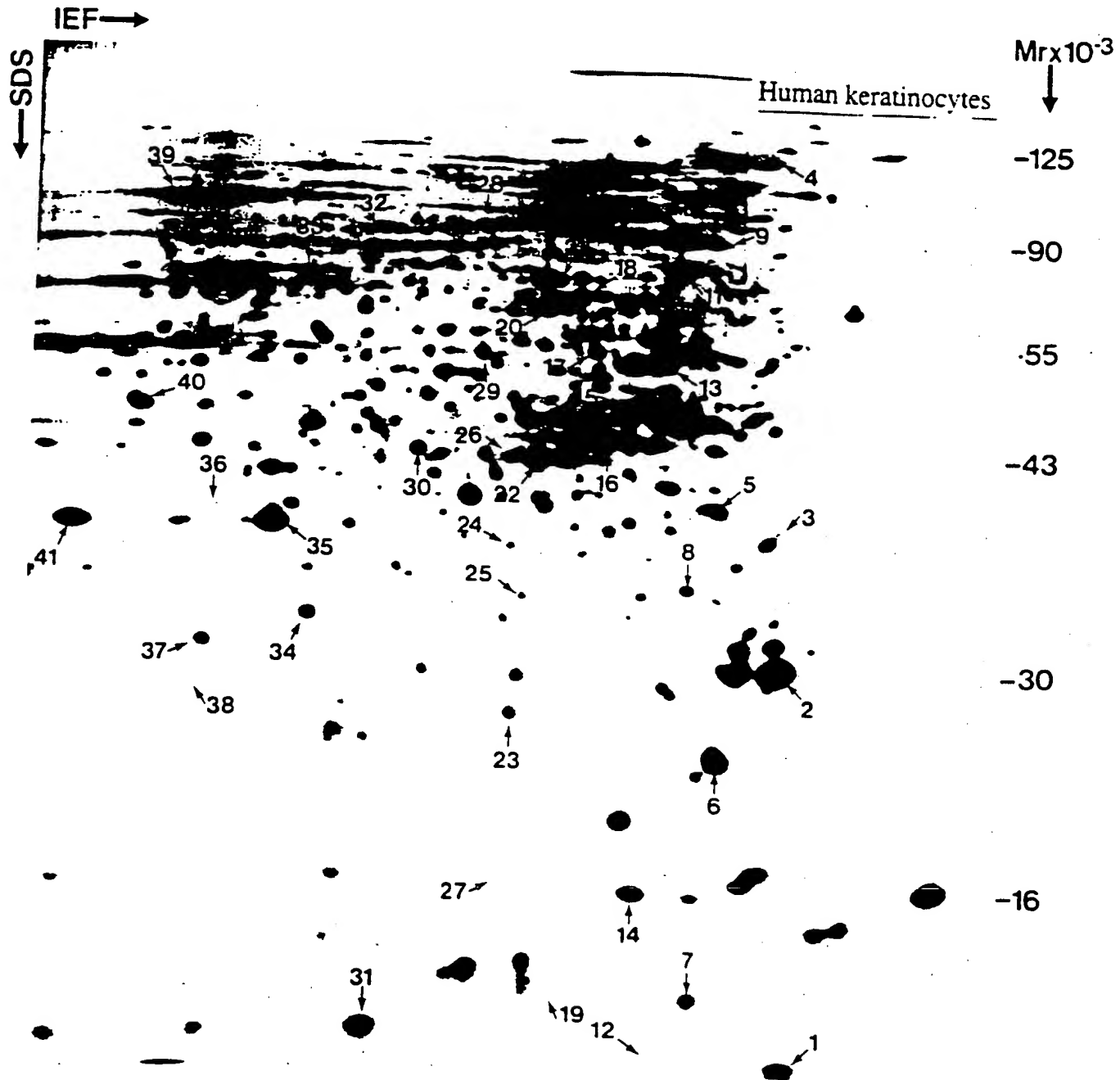


Figure 2. 2-D gel protein map of [³⁵S]methionine-labeled proteins from noncultured, unfractionated normal human keratinocytes focused with CA-IEF in the first dimension. The position of the 41 proteins analyzed in this study is indicated.

Table 2. Proteins from the human keratinocyte database localized in 2-D gels run with IPGs as first dimension

Number in Figs. 1-3	Protein name	IP-1 SSP number ^{a)}	Experimental pI value	Calculated pI value	Discrepancy (pI units)	Calculated net charge at experimental pI value	Buffer capacity charge units per pI unit	N-terminal residues	Recalculated for suspected blockage	Discrepancy pI units	Net charge	Swiss-Prot accession number
1	CaN 19	9027	4.46									
2	Stratidin, bovine 14.3.3 related protein	9109	4.58									
3	Proliferating nuclear antigen (PCNA)/cyclin	9226	4.58									
4	Involucrin	9703	4.63	4.57	-0.01	-0.1	20.8	M				P12004
5	Nucleolar protein B23	8207	4.75	4.63	-0.01	-0.3	70.1	M				P07476
6	Transcriptionally controlled tumor protein	8114	4.79	4.64	-0.11	-3.2	30.4	M				P06748
7	Thioredoxin	8006	4.86	4.84	-0.05	0.6	13.1	M ⁹⁰				P13693
8	Annexin V	8213	4.89	4.82	-0.04	0.3	7.1	V ⁹⁰				P10509
9	Heat shock protein 90- β	8611	4.95	4.88	-0.01	0.1	20.3	A ⁹⁰				P08758
10	Heat shock protein 90- α	2629	4.97	4.94	-0.01	-0.5	56.2	P				P07900
11	Glucose regulated protein 78 (Grp)	8515	4.99	4.97	0.00	0.2	53.6	P				P08218
12	Calnexin	8017	5.02	4.98	-0.01	-0.6	37.5	E				P11021
13	Vimentin	8117	5.05	5.32	0.30	1.3	3.6	M				P06703
14	Initiation factor 4D	8016	5.05	5.06	0.01	0.2	27.1	S				P08670
15	Keratin 14	7305	5.08	5.08	0.03	0.2	7.6	A ⁹⁰				P10159
16	β -Actin	7316	5.21	5.21	0.00	0.2	21.0	T				P02533
17	Heat shock protein 60	6403	5.23	5.24	0.01	0.1	13.3	D ⁹⁰				P02570
18	Heat shock cognate 71kD	6501	5.28	5.37	0.09	1.8	17.5	A ⁹⁰				P10809
19	Cystatin	6011	5.30	5.37	0.08	0.2	18.1	M				P11142
20	I-plasin	6112	5.31	5.31	0.07	1.1	3.0	M				P01040
21	Calnexin	5628	5.35	5.37	0.02	0.5	23.3	A ⁹⁰				P13797
22	Plasminogen activator inhibitor-2	6314	5.38	5.36	-0.02	0.9	10.7	M				P08133
23	Glutathione S-transferase 3	5101	5.43	5.44	0.01	0.08	3.9	P				P05120
24	Annexin VIII	5211	5.45	5.46	0.01	1.0	8.7	M				P09211
25	Annexin III	5204	5.46	5.63	0.17	1.4	8.4	M				P13928
26	Adenosine deaminase	5305	5.47	5.63	0.16	1.8	10.8	M				P12429
27	Stathmin	5001	5.55	5.61	0.06	0.4	6.6	A ⁹⁰				P00813
28	Gelsolin, cytoplasmic	5608	5.59	5.58	-0.01	-0.1	16.5	V				P16949
29	Rat phosphomuscle specific protein homolog	5110	5.62									P06396
30	Elastase inhibitor	4314	5.74									
31	ST00, calgizarin	4006	5.75									
32	Cytodifin, ezrin	3504	5.99	5.95	-0.04	0.5	13.2	P				P15111
33	Moecin	3515	6.11	6.09	-0.02	0.2	9.8	P				P26038
34	Purine nucleoside phosphorylase	2108	6.11	6.15	0.34	1.8	1.1	M				P00491
35	Annexin I	2216	6.18	6.64	0.46	1.6	2.5	A				P01081
36	Aldose reductase	1202	6.40	6.55	0.15	0.7	4.2	A				P15121
37	Phosphoglycerate mutase (B form)	1107	6.46	6.75	0.29	0.9	2.6	A				P18669
38	Trisphosphate isomerase	1111	6.53	6.51	-0.02	-0.04	2.3	A ⁹⁰				P00918
39	Elongation factor 2	1610	6.43	6.38	-0.05	0.5	9.8	M				P16109
40	α -Enolase	1325	6.62	6.99	0.37	1.0	2.2	S				P06711
41	Annexin II	210	7.30	7.36	0.06	0.05	0.9	S ⁹⁰				P07155

a) SSP number in the keratinocyte database [15]

b) Peptides N-terminally sequenced as liver proteins [3]

c) Peptides given as N-terminally blocked in Swiss-Prot database

3.2 Comparison between the determined and calculated pI values for human keratinocyte proteins

Thirty six of the 41 proteins listed in Table 2 are found in the Swiss-Prot database. Contrary to the plasma and liver proteins used in [9], the pI calculations on the proteins used in this study posed some problems that reflected the way in which they were characterized. The

proteins used by Bjellqvist *et al.* [9] were either very abundant and well-characterized plasma proteins or they were identified by N-terminal sequencing and, therefore, the nature of the N-terminals (acetylated or non-acetylated) was in both cases known. The proteins used in this study have all been characterized by internal sequencing [7] and it is known that N-terminal acetylation occurs with high frequency in eukaryotes.

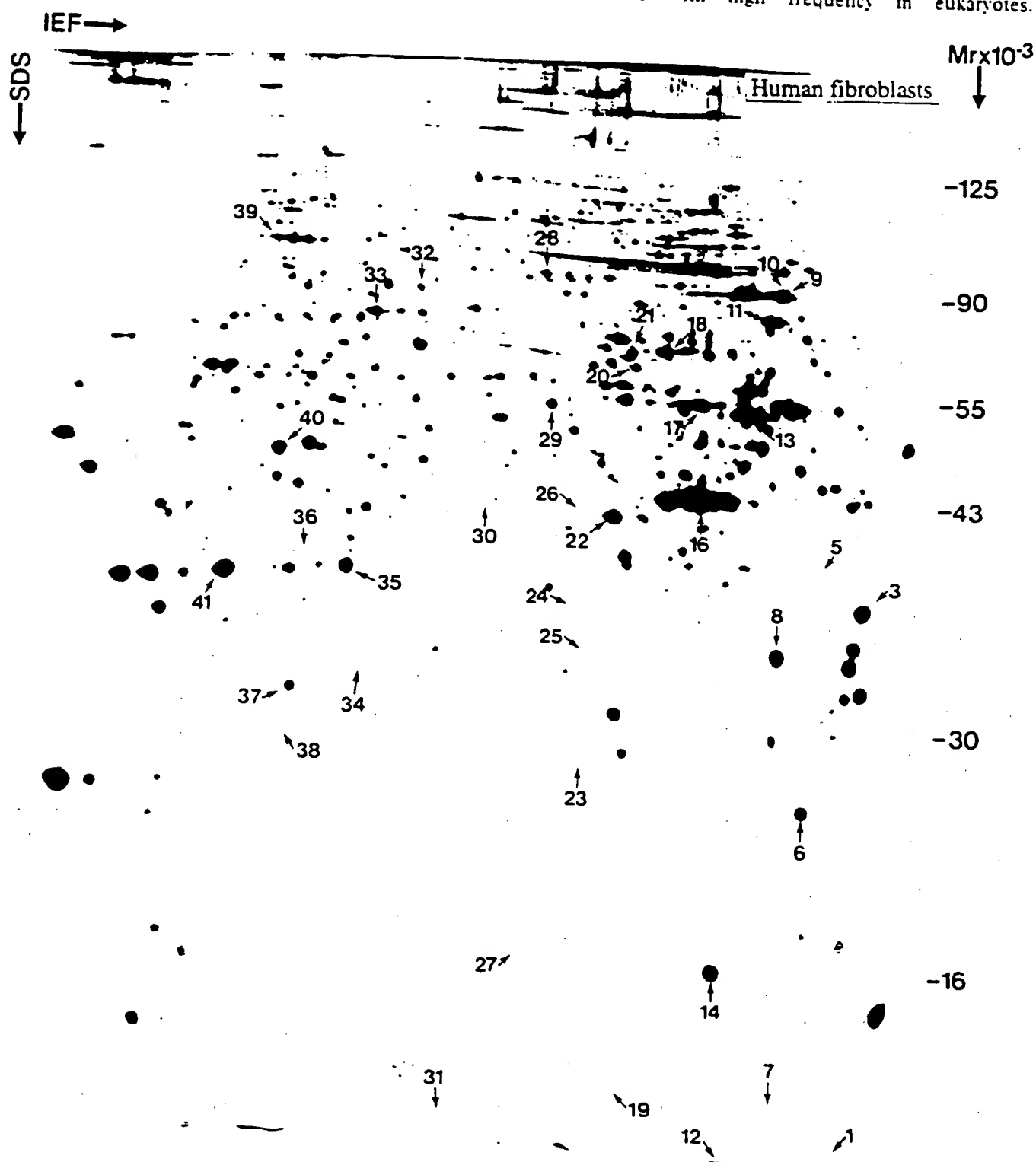


Figure 2. 2-D protein map of [35 S]methionine-labeled proteins from normal human fibroblasts focused with the nonlinear, wide-range IPG in the first dimension. The position of the 41 proteins analyzed in this study is indicated.

According to Brown and Robert [25], proteins with acetylated *N*-terminals correspond in weight to approximately 80% of the soluble protein in ascites cells. Based on results from *N*-terminal sequencing, at least 40% of the spots in the human liver protein 2-D gel map appear to be blocked [3]. The corresponding number, derived from 107 spots in the 2-D gel map of human T-lymphocyte proteins, falls between 60 and 65% (J. Strahler, personal communication). Information concerning *N*-terminal blockage is not normally available, and in the Swiss-Prot database only 6 of the 36 keratinocyte proteins are specified as *N*-terminally blocked. We have, within the present material, defined 18 proteins for which the *N*-terminals are very likely to be correctly described. Six of these proteins are listed in the Swiss-Prot database as *N*-terminally blocked, four represent proteins which appear in the human liver 2-D gel map and have been *N*-terminally sequenced as liver proteins [3] and the remaining eight have *N*-terminal groups other than M, S and A, *i.e.* *N*-terminals for which *N*-acetylation is uncommon [26]. In Figs. 4A, B, C and D *pI* values calculated from Swiss Prot database information are plotted against the experi-

mentally determined *pI* values for all the keratinocyte proteins listed in Table 2 and for the 18 selected proteins, as well as for the plasma and liver proteins (data from [9] valid for 10°C)*.

The calculations show that without knowledge of the status of the *N*-terminal group, precise predictions of *pI* values for eukaryotic proteins cannot be achieved based on the information available in Swiss-Prot and similar databases. However, for proteins where the *N*-terminal status is known, we find good correlation between predicted and experimental *pI* values. When the variance of the *pI* discrepancies and the variance of calculated charges at the experimental *pI* values derived from the present data set are compared with the corresponding

* There are four plots: (A) the 36 polypeptides from normal human keratinocytes (no corrections), (B) the 36 polypeptides from Fig. 4A where *pI* values have been recalculated for 12 polypeptides with M, S and A as *N*-terminally assumed blocked, based on calculated charge, (C) the 18 selected polypeptides with information on the *N*-terminal configuration, and (D) plasma and liver proteins.

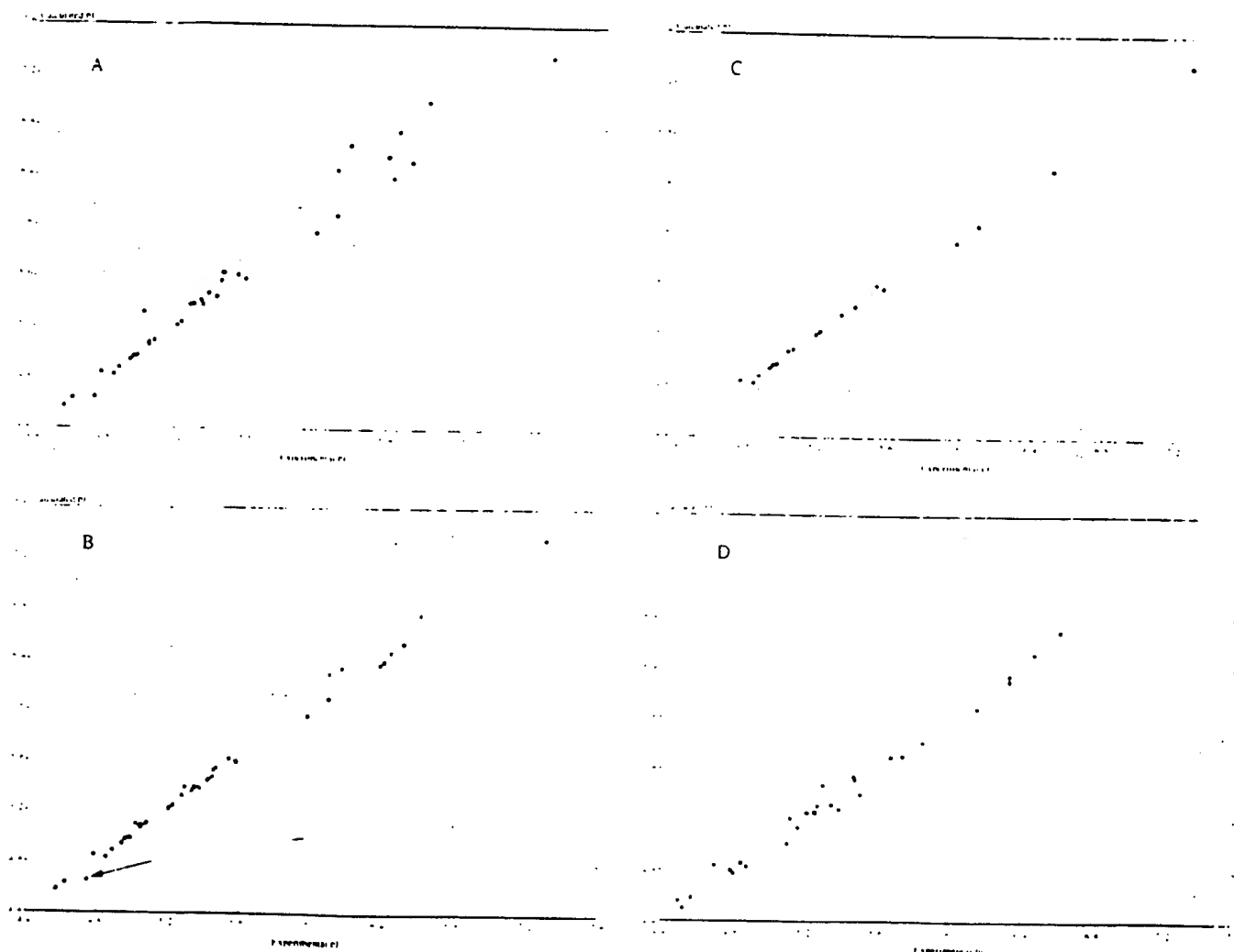


Figure 4. Calculated vs. experimental *pI* values. Lines are fitted using the least squares' criterion. (A) 36 polypeptides from normal human keratinocytes (no corrections). (B) 36 polypeptides from Fig. 4A (including the 18 marker polypeptides) where *pI* values have been recalculated assuming *N*-terminal blockage; x indicates recalculated *pI* values; nucleolar protein B23 is indicated with an arrow. (C) 18 polypeptides with information on *N*-terminal configuration and (D) plasma and liver proteins.

values derived from the data on plasma and liver proteins in [9] (Table 3), the present data are found to result in larger variances for the values of both pI discrepancies and calculated charge at the experimental pI value when no information on posttranslational modification is taken into consideration. Correction for possible N -acetylation of 12 polypeptides with M, S and A as N -terminal results in a smaller variance of pI discrepancies, although not significantly different from values derived from [9], whereas the variance of the calculated charge at the experimental pI value is significantly higher. For the 18 selected proteins the variance for the pI discrepancies is significantly smaller than for the data in [9]; however, the corresponding value for calculated charge at the experimental pI value does not improve to the same extent. This, we believe, reflects another difference between the two sets of proteins used for the calculations. Based on spot distributions in 2-D gel maps, the set of proteins used here has a molecular weight distribution that is more representative of the patterns observed in mammalian cells. In the study by Bjellqvist *et al.* [9] most of the high molecular weight plasma proteins had to be excluded due to their unknown content of sialic acid which made the proteins analyzed in this study heavily biased towards low molecular weight proteins. The buffer capacity of proteins normally increases with the protein's molecular weight, and the average buffer capacity of the presently selected proteins with assumed known N -terminals is 18 charge units/pH unit, while the corresponding value for the proteins used in [9] is only 9 charge units/pH unit. High buffer capacity can be expected to improve the agreement between calculated and experimental pI values. Inspection of the data presented in Table 2 for the polypeptides with assumed known N -terminals verifies the importance of the buffer capacity. For 8 polypeptides having buffer capacities higher than 15 charge units/pH unit, the calculations in all cases yielded pI discrepancies with absolute values of less than 0.02 pH units. The largest discrepancy, 0.06 pH units, was observed for annexin II and stathmin, proteins which have low buffer capacity: 0.9

and 6.6 charge units/pH unit, respectively. The probability that the focusing position of a protein with known composition will fall within a certain distance from the calculated pI value therefore cannot be predicted by the variance alone. The buffer capacity of the specific protein must be taken into consideration as well. As indicated by the decrease of the variance of calculated charges at the experimental pI value for the selected proteins, the observed improvement can not solely be due to the higher buffer capacity of the keratinocyte proteins. The two studies relate to different experimental conditions. Good agreement between experimental and calculated pI values implies that the proteins are defolded and a factor that may contribute to the observed improvement is a more complete defolding of proteins caused by the higher temperature and urea concentration used in this study.

The data indicated that the precision with which pI values can be predicted for polypeptides with high buffer capacity is better than the precision with which experimental pI values can be determined. If the pH is defined through the pK values of the immobilized groups in the IPG containing gel, the precision of the experimentally calculated data will depend on the pH difference between the pI and the pK value of the immobilized group with the closest pK . For the present study this will give pI determinations with a precision varying in the range of ± 0.02 – 0.05 pH units [9]. The good agreement observed between the calculated and experimental pI values is due to the fact that errors are mainly systematic and, as discussed in [9], they will largely be cancelled out in the calculations. A pH scale defined through the presently determined pI values will not necessarily reflect the variation of the hydrogen ion activity during the focusing step in an optimal way, but it still allows precise predictions of focusing positions for polypeptides with known compositions, including information on posttranslational modifications. Calculated net charge at the experimentally found isoelectric point defined in this scale will serve as a tool to verify that the polypeptide

Table 3. Mean values and variances for the difference (experimental pI -calculated pI) in pH units and calculated charges at the experimental pI values, respectively

Number of proteins	Plasma and liver proteins (8 M urea, 10°C)				Keratinocyte proteins (9.6 M urea, 25°C)			
			All peptides		All peptides after correction for N -acetylation		Known N -terminal configuration (or very likely configuration)	
	29		36		36		18	
Experimental pI -calculated pI	Mean -0.011	Variance 0.005	Mean 0.072	Variance 0.017	Mean 0.019	Variance 0.003	Mean 0.005	Variance 0.001
F-value (pI discrepancy) ¹⁾	1		3.4		1.67		5	
P-level (pI discrepancy) ²⁾	0.5		0.0005		0.0721		0.0004	
Calculated charge at the experimental pI value	-0.070	0.22	0.321	0.871	0.009	0.444	-0.014	0.109
F-value (calculated charge at the experimental pI value) ¹⁾	1		3.8		1.96		2.08	
P-level (calculated charge at the experimental pI value) ²⁾	0.5		0.0002		0.0338		0.0536	

1) Comparison to the data in [9]. $F = S_1^2/S_2^2$, where S_1^2 is the larger of the two variances

2) $P(F(v_1, v_2) \geq F\text{-value})$, where v_1 and v_2 are the degrees of freedom for s_1 and s_2 , respectively

composition used in the calculation is correct and complete. Exceptions to this are proteins such as involucrin and heat shock protein 90 that have very high buffer capacities. Introduction of an extra charge unit into these proteins will only result in pI shifts falling in the range of 0.01–0.02 pH units and the effect is that the quality of the pH definition – the precision by which pK values used in the calculations are given and the precision of experimental pI values in these cases – will limit the possibilities to verify polypeptide composition based on the experimental pI value.

Statistical comparison of experimental and calculated pI values was done using the t -test for dependent samples and normality of the discrepancies was estimated by probability plots. For the 36 proteins, the p -level is 0.0021, indicating that a result like this is unlikely to be a chance effect and must be assumed to represent a real difference. After correction for the most likely N -terminal configuration, the p -level is 0.043 and cannot be accepted as representing the same population since the p -level is less than 0.05 – the traditional p -limit of statistical significance. For the 18 proteins with a known or very likely N -terminal configuration the t -test gave a p -level of 0.49, which verifies that the experimental and calculated pI values are not significantly different.

Besides showing that pI values for denatured proteins with known compositions can be calculated with a high degree of precision from average pK values, the results also provide strong support for the notion that N -terminal blockage heavily depends on the nature of the N -terminal groups [26]. The results seem to indicate that with N -terminals other than M, S and A, only a few proteins have blocked N -terminals (1 out of 10 proteins in the present study), while it can be inferred from the data presented in Table 2 that a majority of the proteins with M, S and A as N -terminal are blocked. After correction for the effect of suspected N -terminal blockage there is only one protein (nucleolar protein B23) out of the 36 used in this study, which, in spite of a high buffer capacity, has a marked difference of 0.11 pH units between predicted and determined pI values (Fig. 4B); this corresponds to 3 charge units due to the high buffer capacity of this protein. This discrepancy in pI prediction and calculation of net charge at the pI is probably not due to deficiencies in the database information but instead reflects a shortcoming of the model used for pI calculations. Nucleolar protein B23 contains a domain extremely rich in aspartic and glutamic acid residues (Table 4), in which 26 out of 28 amino acid residues from position 161 to 188 are either a D or an E. A calculation based on the use of average pK values uninfluenced by the charged neighboring amino acid residues cannot be expected to correctly describe the pI value with almost half of the acidic groups packed

together into a highly negatively charged region. This limitation caused by calculations based on average pK values does not severely limit the usefulness of the approach since a search through Swiss-Prot shows that this type of D/E-rich motif is uncommon, and the existence of a highly charged region is immediately apparent upon inspection of the amino acid sequence.

The quality of the information available in databases, especially concerning posttranslational modifications, is a major problem when the data is to be used for pI predictions. The p -level of 0.043 found for all 36 proteins after correction for N -acetylation, shows that this problem is not only limited to N -terminal blockage and the very good agreement found for the eighteen polypeptides, with assumingly correctly described N -terminal (Fig. 4C), must be regarded as an exception from this point of view. N -Terminal blockage is generally the main problem in relation to pI predictions for eukaryotic proteins. Of the 36 keratinocyte proteins analyzed, 18–20 are suspected to be N -terminally blocked (6 proteins blocked according to Swiss-Prot, 12 proteins with M, S or A as N -terminal and assumingly blocked based on the calculated charge, and two proteins, involucrin and nucleolar protein B23, with M as N -terminal for which the data does not allow any conclusion). This is in reasonable agreement with the conclusions based on the N -terminal sequencing data derived in connection with 2-D gel electrophoresis. N -terminal blockage can be suspected for 17–19 of the 26 proteins with M, S or A as N -terminal, while only 1 in 10 proteins with other N -terminal groups are blocked. The information that the frequency of N -terminal blockage is strongly related to the nature of the N -terminal group will be of some help in connection with pI predictions based on database information. However, without information from other sources, an uncertainty will always remain as to whether the N -terminal charge should be included in the pI calculation.

4 Concluding remarks

The data presented here lays the foundation for comparing 2-D gel protein maps of different cell types generated with nonlinear, wide-range IPGs in the first dimension. The focusing positions of 41 polypeptides common to most human cell types have been described in a pH scale that allows focusing positions to be predicted with a high degree of accuracy, provided that the composition of the polypeptides are known and that information on posttranslational modifications are available. For polypeptides with a very high buffer capacity, the limiting factor is the precision with which experimental pH values can be determined rather than the precision of the calculations. Possible deficiencies in the pH scale description of the variation of the hydrogen ion activity has, at least at the present state, no consequences for its practical use. The major limitation in connection with predictions of focusing positions from polypeptide compositions is the quality of existing data on protein compositions, especially concerning posttranslational modifications. Amino acid sequences have been reasonably easy to obtain, while posttranslational modifications

Table 4. Amino acid sequence of nucleolar phosphoprotein B23

1	MESSDQDS	FLPPQALPS	TELVADQTH	FRVQDERH	QLSQTHVDS
51	AGVDELTH	EASQKQSS	FRVTLATL	KNVPTVDS	QFTETPAV
101	FLNGSSPFT	LSQKQKVE	ESASQSSD	ETVLLVDS	FRASQSSD
151	TPQKATLA	TEVQKQSS	FRVQDERH	FRVQDERH	FRVQDERH
201	APQKQKTH	FRVQDERH	FRVQDERH	FRVQDERH	FRVQDERH
251	MSQKQSS	FRVQDERH	FRVQDERH	FRVQDERH	FRVQDERH

have been difficult and work-intensive to determine. Recent developments in the field of mass spectrometry are fast changing this situation and within the next years we can expect a surge in reliable data in this area. While awaiting this development, verification of correctness and completeness of available information on polypeptide composition can be provided by experimental pI values in a pH scale based on the pI values determined in this study. So far, our data cover the pH range below $pH \approx 7.5$. The basic pH range covered by NEPHGE as first dimension will be covered in forthcoming work.

Received December 29, 1993

5 References

- [1] Gianazza, E., Astrua-Testori, S., Caccia, P., Giaccon, P., Quaglia, L., Righetti, P. G., *Electrophoresis* 1986, 7, 76-83.
- [2] Görg, A., Postel, W., Günther, S., *Electrophoresis* 1988, 9, 531-546.
- [3] Hochstrasser, D. F., Frutiger, S., Paquet, N., Bairoch, A., Ravier, F., Pasquali, C., Sanchez, J.-C., Tissot, J.-D., Bjellqvist, B., Vargas, R., Appel, R. D., Hughes, G. J., *Electrophoresis* 1992, 13, 992-1001.
- [4] *Immobiline Dr-Strip Kit for 2-D Electrophoresis: Instructions*, Pharmacia LKB Biotechnology AB, Uppsala 1993.
- [5] Anderson, N. L., Hickman, B. J., *Anal. Biochem.* 1979, 93, 312-320.
- [6] Neidhardt, F. C., Appieby, D. A., Sankar, P., Hutton, M. E., Phillips, T. A., *Electrophoresis* 1989, 10, 116-121.
- [7] Rasmussen, H. H., Damme, J. V., Puype, M., Gesser, B., Celis, J. E., Vandekerckhove, J., *Electrophoresis* 1992, 13, 960-969.
- [8] Gianazza, E., Artoni, G., Righetti, P. G., *Electrophoresis* 1983, 4, 321-326.
- [9] Bjellqvist, B., Hughes, G. J., Pasquali, C., Paquet, N., Ravier, F., Sanchez, J.-C., Frutiger, S., Hochstrasser, D. F., *Electrophoresis* 1993, 14, 1023-1031.
- [10] Bjellqvist, B., Pasquali, C., Ravier, C., Sanchez, J.-C., Hochstrasser, D. F., *Electrophoresis* 1993, 14, 1357-1365.
- [11] O'Farrell, P. H., *J. Biol. Chem.* 1975, 250, 4007-4021.
- [12] Görg, A., *Biochem. Soc. Transactions* 1993, 21, 130-132.
- [13] Hanash, S. M., Strahler, J. R., Neel, J. V., Hailat, N., Mathem, R., Keim, D., Zhu, X. X., Wagner, D., Gage, D. A., Watson, J. T., *Proc. Natl. Acad. Sci. USA* 1991, 88, 5709-5713.
- [14] Görg, A., Postel, W., Friedrich, C., Kuck, R., Strahler, J. R., Hanash, S. M., *Electrophoresis* 1991, 12, 653-658.
- [15] Celis, J. E., Rasmussen, H. H., Olsen, E., Madsen, P., Leffers, H., Honore, B., Dejgaard, K., Gromov, P., Hoffmann, H. J., Nielsen, M., Vassilev, A., Vintermyr, O., Hao, J., Celis, A., Basse, B., Lauridsen, J. B., Ratz, G. P., Andersen, A. H., Walbum, E., Kiergaard, L., Puype, M., Van Damme, J., Delay, B., Vandekerckhove, J., *Electrophoresis* 1993, 14, 1091-1198.
- [16] Celis, J. E., Madsen, P., Rasmussen, H. H., Leffers, H., Honore, B., Gesser, B., Dejgaard, K., Olsen, E., Magnusson, N., Kjil, J., Celis, A., Lauridsen, J. B., Basse, B., Ratz, G. P., Andersen, A., Walbum, E., Brandstrup, B., Pedersen, P. S., Brandt, N. J., Puype, M., Van Damme, J., Vandekerckhove, J., *Electrophoresis* 1991, 12, 802-872.
- [17] Bjellqvist, B., Ek, K., Righetti, P. G., Gianazza, E., Görg, A., Postel, W., Westermeier, R., *J. Biochem. Biophys. Methods* 1982, 6, 317-333.
- [18] Bairoch, A., Boeckman, B., *Nucleic Acids Res.* 1991, 19, 2247-2249.
- [19] Honore, B., Madsen, P., Basse, B., Andersen, A., Walbum, E., Celis, J. E., Leffers, H., *Nucleic Acids Res.* 1990, 18, 6692.
- [20] Altland, K., *Electrophoresis* 1990, 11, 140-147.
- [21] Perrin, D. D., Dempsey, B., Serjant, E. P., *pKa Predictions for Organic Acids and Bases*, Chapman and Hall Ltd., London 1981.
- [22] Perrin, D. D., *Dissociation Constants of Organic Bases in Aqueous Solutions*, Butterworths, London 1965.
- [23] Perrin, D. D., *Dissociation Constants of Organic Bases in Aqueous Solutions*, Supplement 1972, Butterworths, London 1972.
- [24] Altland, K., Becher, P., Rossman, U., Bjellqvist, B., *Electrophoresis* 1988, 9, 474-485.
- [25] Brown, J. L., Robert, W. K., *J. Biol. Chem.* 1976, 251, 1009-1014.
- [26] Persson, B., Flinta, C., Heine, G., Jorvall, H., *Eur. J. Biochem.* 1985, 152, 523-527.

Company Info

LSB & LSP Information

Large Scale Biology Corporation

Large Scale Proteomics Corporation

Large Scale Biology Corporation

Large Scale Biology Corporation is the leader in the integrated discovery, production and application of proteins - the functional units of all biological processes.

Large Scale Biology Corporation (LSB, Vacaville, CA) and its subsidiary Large Scale Proteomics Corp. (LSP, Germantown, MD) are a biotechnology enterprise with the mission of accelerating the speed and productivity of the life sciences industry product discovery and development programs. Unique among biotechnology companies is LSB's integration of technologies to discover, analyze, manufacture and find new applications for proteins - the functional units of all biological processes.

Genomics companies have focused on deciphering genetic information, providing an initial but only partial understanding of biological processes. LSB's proprietary protein technologies can enable the transformation of genomic information into products such as drug targets, therapeutics, diagnostics for drug efficacy and toxicity, and traits for agricultural crops. Large Scale Biology has gone beyond the "genomics" realm in its business model and developed ways to integrate the discovery of gene function with quantitative protein analysis and protein manufacturing. This integration of technology platforms favorably positions LSB as a leading provider of valuable content to industry leaders in the fields of diagnostics, therapeutics, vaccines and agribusiness.

LSB was founded in 1987 with the goal of commercializing its proprietary GENEWARE viral vector system - a novel technology for gene expression. Using safe RNA viruses to transiently express genes in non-recombinant plants, LSB has positioned itself in the industry to provide cost-effective manufacturing and purification of diverse protein and peptide products. The same technology can be applied to the expression of libraries of foreign genes in an automated, high-throughput format to discover the function of genes with unparalleled efficiency. The GENEWARE system and associated proprietary technologies form the basis for LSB's functional genomics, biomanufacturing and a variety of proprietary products under development.

From its foundation, LSB understood the need to integrate functional genomic and protein manufacturing expertise with quantitative protein analysis and informatics to become a world-leader in the protein field. In 1999, LSB acquired a privately held pharmaceutical proteomics company originally founded in 1985. Large Scale Proteomics Corporation (a wholly

owned subsidiary of Large Scale Biology Corporation) is an industry leader in identifying and characterizing proteins in all types of biological samples for the discovery and development of new and more effective therapies, diagnostics, and agricultural products.

"Proteomics" is the study of the entire complement of proteins expressed in a cell, tissue, or organism. Proteomics can significantly improve drug discovery and development because most illness is associated with imbalances among, or malfunctions of, proteins. Only a small fraction of diseases can be attributed to the presence of a defective gene. Unlike classical genomics approaches that discover genes that may relate to a disease, LSP has developed a proprietary system called the ProGEx module for directly characterizing proteins associated with disease. Using this same technology, LSP can characterize the effects of candidate drugs intended to reverse a disease process, and to determine the degree to which this objective is achieved free of adverse side effects.

LSB and LSP have protected their many discoveries through an extensive portfolio of domestic and foreign patents and have developed commercial alliances and partnerships to exploit the value of their technologies. LSB and LSP scientists and engineers focus on the development and application of resources to help clients meet their objectives as well as the development of our own proprietary products for subsequent partnering with industry leaders.

A combined staff of 140 professionals operates from three locations in the United States, with a network of collaborators and affiliates throughout the US and Europe. Company headquarters, R&D laboratories and its Genomics division are located in Vacaville, California about 60 miles northeast of San Francisco. Process development and biomanufacturing take place in Owensboro, Kentucky, and LSB's Large Scale Proteomics Corporation subsidiary is located in Germantown, Maryland.

In August, 2000, LSB completed an initial public offering (IPO) of 5 million shares of common stock and now trades on the NASDAQ under the symbol LSBC.

Leadership - Large Scale Biology Corporation

Robert L. Erwin, Chairman of the Board and Chief Executive Officer, founded LSB™ and has served as a director and officer since 1987. Mr. Erwin is the former chairman of the State of California Breast Cancer Research Council and currently serves on the University of California President's Engineering Advisory Council. He is Chairman of the Supervisory Board of Icon Genetics AG. As a co-founder of Sungene Technologies Corp., Mr. Erwin served as Vice President of Research and Product Development from 1981 through 1986. He has served on the Biotechnology Industry Advisory Board for Iowa State University. Mr. Erwin received his M.S. degree in Genetics from Louisiana State University and is an inventor on several LSB patents.

David R. McGee, Ph.D., a co-founder of LSB and Senior Vice President and Chief Operating Officer, has been an officer since 1987. Prior to joining LSB, Dr. McGee was Vice President of Operations at Sungene Technologies Corporation from 1983 to 1987. Dr. McGee received his Ph.D. in Genetics from Louisiana State University and served as a faculty instructor of zoology and genetics at Louisiana State University.

Laurence K. Grill, Ph.D., a co-founder of LSB and Senior Vice President, Research and Development, has served as an officer since 1987. Dr. Grill was the Manager of Plant Molecular Biology for Sandoz Crop Protection Corp. from 1984 to 1987 and Senior Research

Scientist in the Department of Molecular Biology at Zoecon Research Institute from 1980 to 1984. He received his Ph.D. from the University of California at Riverside with an emphasis on the molecular basis for viral gene expression in plants.

R. Barry Holtz, Ph. D., Senior Vice President, Biopharmaceutical Manufacturing, has served the company as an officer since 1989 upon the acquisition of Holtz Bio-Engineering, which was founded in 1980. Dr. Holtz was a co-founder and Director of Research for MFI, Inc., the largest manufacturer of microencapsulated nutrients for agriculture and Director of Fundamental Research at Foremost-McKesson, Inc. Dr. Holtz received his Ph.D. in Biochemistry from Pennsylvania State University and served as Assistant Professor in the Department of Food Science and Nutrition at Ohio State University.

Daniel Tusé, Ph.D., has been an officer of LSB since he joined the Company in 1995 as Vice President, Pharmaceutical Development. Dr. Tusé manages the company's pharmaceutical design and development programs, including LSB's novel vaccines and immunotherapeutics initiatives. Prior to joining LSB, Dr. Tusé was Assistant Director of SRI International's (Menlo Park, Calif.) Life Sciences Division. In his 17 years at SRI, Dr. Tusé developed extensive R&D experience in pharmaceuticals and specialty chemicals, serving an international list of clients. Dr. Tusé received his Ph.D. in Microbiology (1980, *cum laude*) with a minor in Toxicology from the University of California, Davis.

John S. Rakitan, a co-founder of LSB, Senior Vice President & General Counsel and Secretary, has served as an officer since 1988. Prior to joining LSB, Mr. Rakitan was an attorney in private practice. Mr. Rakitan received his J.D. degree from the University of Notre Dame.

Michael D. Centron, Treasurer, has served as Controller since 1988 and was elected as Treasurer in 1991. Mr. Centron was Audit Supervisor for Varian Associates from June 1985 through July 1988, and he also worked for Arthur Young and Co. (currently Ernst & Young). Mr. Centron is a certified public accountant and received his M.B.A. degree from the University of California at Berkeley.

Guy della-Cioppa, Ph.D., is an officer of the company and currently serves as Vice President, Genomics. Prior to joining the company in 1989, Dr. della-Cioppa worked for Monsanto Company in St. Louis, MO from 1984-1989 and was an NIH Postdoctoral Fellow at the Worcester Foundation for Experimental Biology in Shrewsbury, MA from 1983-1984. He received his Ph.D. in Biology from the University of California, Los Angeles.

William M. Pfann joined Large Scale Biology in August 2000 as Senior Vice President Finance and Chief Financial Officer. Mr. Pfann was formerly with PricewaterhouseCoopers LLP from 1969 to July 2000, most recently as the Risk Management Partner for the Western Region. He served in a number of management roles at PwC, including leader of the firm's Silicon Valley audit practice, National Director of the networking and communications sector and Managing Partner of the Northern California emerging business group, as well as Partner-in-Charge of the Oakland and Walnut Creek, California offices. Mr. Pfann received a B.S. degree from the University of California, Berkeley, in Business Administration and an MBA in Accounting from Golden Gate University.

[back to index](#)

Large Scale Proteomics Corporation

Leadership - Large Scale Proteomics Corporation

N. Leigh Anderson, Ph.D., Chairman, President and CEO of Large Scale Proteomics Corporation (LSP™). Dr. Anderson obtained his B.A. in Physics with honors from Yale and a Ph.D. in Molecular Biology from Cambridge University (England) working with M. F. Perutz as a Churchill Fellow at the MRC Laboratory of Molecular Biology. Subsequently he co-founded the Molecular Anatomy Program at the Argonne National Laboratory (Chicago) where his work in the development of 2-dimensional electrophoresis (2-DE) and molecular database technology earned him, among other distinctions, the American Association for Clinical Chemistry's Young Investigator Award for 1982 and the 1983 Pittsburgh Analytical Chemistry Award. In 1985 Dr. Anderson co-founded LSP (originally Large Scale Biology Corp., Germantown, MD) in order to pursue commercial development and large-scale applications of 2-D electrophoretic protein mapping technology.

Norman G. Anderson, Ph.D., Chief Scientist at LSP. Dr. Anderson has a distinguished record as an inventor. His career includes senior positions at Oak Ridge and Argonne National Laboratories (ORNL and ANL), more than 300 scientific publications, and the receipt of more than 20 prestigious awards in recognition of his work in science and technology. For his invention of the zonal ultracentrifuge, he received the John Scott Medal Award, and for the centrifugal fast analyzer, the Preis Biochemische Analytik für Klinische Chemie from Die Deutsche Gesellschaft für Klinische Chemie for the most outstanding analytical development in clinical chemistry worldwide during a 2-year period. In 1984 ANL awarded him its career patent leader award for the largest number of patents issued to an employee. At that time the commercial value of his inventions in terms of U.S. sales and royalties from foreign licensing were \$250 million and \$1 million, respectively. Dr. Anderson received his degrees at Duke University: a B.A. in Zoology, M.A. in Physiology, and Ph.D. in Cell Physiology. He holds 28 patents.

Constance Seniff, Vice President, Operations. Ms. Seniff has managed LSP's operations since 1993. Her background includes thirteen years in international business prior to joining LSP, five abroad in the employ of foreign firms. Ms. Seniff is responsible for helping formulate and implement business development and database commercialization strategies for LSP in coordination with the management of LSP's parent company, Large Scale Biology Corporation. Ms. Seniff has a B.Sc. degree in Business (with honors) from Florida State University.

Robert J. Walden, Vice President, Finance at LSP. Mr. Walden joined LSP in 1997 and has served as a director since 1999. He previously served as Vice President of Finance and Administration at Osiris Therapeutics, Inc., and as Chief Financial Officer at the American Type Culture Collection (ATCC). Mr. Walden received his degree in Finance from the University of Maryland.

Jean-Paul Hofmann, Ph.D., Vice President, Software Development at LSP. Dr. Hofmann is a plant geneticist by training, having earned a B.S. in Biology, M.S. in Biochemistry and Genetics, and Ph.D. in Plant Genetics from the University of Orsay, Paris. He has extensive

experience in using 2-DE in agronomic research and in designing analytical software for 1- and 2-D applications. He has held senior scientific positions in industry and research institutes, in the U.S., France and the Ivory Coast.

John Taylor, Ph.D., Vice President, Software Development and Bioinformatics. Dr. Taylor is the principal developer of Kepler™, LSP's analytical software for automated 2-DE pattern analysis. Prior to joining LSB, Dr. Taylor served as computer scientist in the Molecular Anatomy Program at Argonne, and on the research staffs of the University of Chicago and the Armed Forces Institute of Pathology in Washington, D.C. Dr. Taylor received a B.S. in Physics from the University of South Carolina, and a Ph.D. in Nuclear Physics from Duke University.

Sandra Steiner, Ph.D., currently serves as Vice President Proteomics Applications. Prior to joining the Company, Dr. Steiner founded and directed the Molecular Toxicology Group at Novartis in Basel, Switzerland and was a member in several multi-disciplinary drug development project teams. Dr. Steiner received her Ph.D. in Toxicology/Pharmacology from the University of Basel, Switzerland.

[back to index](#)

© 2000 Large Scale Biology Corporation. All Rights Reserved Worldwide.

Microarrays and Toxicology: The Advent of Toxicogenomics

Emile F. Nuwaysir,¹ Michael Bittner,² Jeffrey Trent,² J. Carl Barrett,¹ and Cynthia A. Afshari¹

¹Laboratory of Molecular Carcinogenesis, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina

²Laboratory of Cancer Genetics, National Human Genome Research Institute, Bethesda, Maryland

The availability of genome-scale DNA sequence information and reagents has radically altered life-science research. This revolution has led to the development of a new scientific subdiscipline derived from a combination of the fields of toxicology and genomics. This subdiscipline, termed toxicogenomics, is concerned with the identification of potential human and environmental toxicants, and their putative mechanisms of action, through the use of genomics resources. One such resource is DNA microarrays or "chips," which allow the monitoring of the expression levels of thousands of genes simultaneously. Here we propose a general method by which gene expression, as measured by cDNA microarrays, can be used as a highly sensitive and informative marker for toxicity. Our purpose is to acquaint the reader with the development and current state of microarray technology and to present our view of the usefulness of microarrays to the field of toxicology. *Mol. Carcinog.* 24:153–159, 1999. © 1999 Wiley-Liss, Inc.

Key words: toxicology; gene expression; animal bioassay

INTRODUCTION

Technological advancements combined with intensive DNA sequencing efforts have generated an enormous database of sequence information over the past decade. To date, more than 3 million sequences, totaling over 2.2 billion bases [1], are contained within the GenBank database, which includes the complete sequences of 19 different organisms [2]. The first complete sequence of a free-living organism, *Haemophilus influenzae*, was reported in 1995 [3] and was followed shortly thereafter by the first complete sequence of a eukaryote, *Saccharomyces cerevisiae* [4]. The development of dramatically improved sequencing methodologies promises that complete elucidation of the *Homo sapiens* DNA sequence is not far behind [5].

To exploit more fully the wealth of new sequence information, it was necessary to develop novel methods for the high-throughput or parallel monitoring of gene expression. Established methods such as northern blotting, RNase protection assays, S1 nuclease analysis, plaque hybridization, and slot blots do not provide sufficient throughput to effectively utilize the new genomics resources. Newer methods such as differential display [6], high-density filter hybridization [7,8], serial analysis of gene expression [9], and cDNA- and oligonucleotide-based microarray "chip" hybridization [10–12] are possible solutions to this bottleneck. It is our belief that the microarray approach, which allows the monitoring of expression levels of thousands of genes simultaneously, is a tool of unprecedented power for use in toxicology studies.

Almost without exception, gene expression is altered during toxicity, as either a direct or indirect result of toxicant exposure. The challenge facing toxicologists is to define, under a given set of experimental conditions, the characteristic and specific pattern of gene expression elicited by a given toxicant. Microarray technology offers an ideal platform for this type of analysis and could be the foundation for a fundamentally new approach to toxicology testing.

MICROARRAY DEVELOPMENT AND APPLICATIONS

cDNA Microarrays

In the past several years, numerous systems were developed for the construction of large-scale DNA arrays. All of these platforms are based on cDNAs or oligonucleotides immobilized to a solid support. In the cDNA approach, cDNA (or genomic) clones of interest are arrayed in a multi-well format and amplified by polymerase chain reaction. The products of this amplification, which are usually 500- to 2000-bp clones from the 3' regions of the genes of interest, are then spotted onto solid support by using high-speed robotics. By using this method, microarrays of up to 10 000 clones can be generated by spotting onto a glass substrate

*Correspondence to: Laboratory of Molecular Carcinogenesis, National Institute of Environmental Health Sciences, 111 Alexander Drive, Research Triangle Park, NC 27709.

Received 8 December 1998; Accepted 5 January 1999

Abbreviations: PAH, polycyclic aromatic hydrocarbon; NIEHS, National Institute of Environmental Health Sciences.

[13,14]. Sample detection for microarrays on glass involves the use of probes labeled with fluorescent or radioactive nucleotides.

Fluorescent cDNA probes are generated from control and test RNA samples in single-round reverse-transcription reactions in the presence of fluorescently tagged dUTP (e.g., Cy3-dUTP and Cy5-dUTP), which produces control and test products labeled with different fluorophores. The cDNAs generated from these two populations, collectively termed the "probe," are then mixed and hybridized to the array under a glass coverslip [10,11,15]. The fluorescent signal is detected by using a custom-designed scanning confocal microscope equipped with a motorized stage and lasers for fluor excitation [10,11,15]. The data are analyzed with custom digital image analysis software that determines for each DNA feature the ratio of fluor 1 to fluor 2, corrected for local background [16,17]. The strength of this approach lies in the ability to label RNAs from control and treated samples with different fluorescent nucleotides, allowing for the simultaneous hybridization and detection of both populations on one microarray. This method eliminates the need to control for hybridization between arrays. The research groups of Drs. Patrick Brown and Ron Davis at Stanford University spearheaded the effort to develop this approach, which has been successfully applied to studies of *Arabidopsis thaliana* RNA [10], yeast genomic DNA [15], tumorigenic versus non-tumorigenic human tumor cell lines [11], human T-cells [18], yeast RNA [19], and human inflammatory disease-related genes [20]. The most dramatic result of this effort was the first published account of gene expression of an entire genome, that of the yeast *Saccharomyces cerevisiae* [21].

In an alternative approach, large numbers of cDNA clones can be spotted onto a membrane support, albeit at a lower density [7,22]. This method is useful for expression profiling and large-scale screening and mapping of genomic or cDNA clones [7,22–24]. In expression profiling on filter membranes, two different membranes are used simultaneously for control and test RNA hybridizations, or a single membrane is stripped and reprobed. The signal is detected by using radioactive nucleotides and visualized by phosphorimager analysis or autoradiography. Numerous companies now sell such cDNA membranes and software to analyze the image data [25–27].

Oligonucleotide Microarrays

Oligonucleotide microarrays are constructed either by spotting prefabricated oligos on a glass support [13] or by the more elegant method of direct in situ oligo synthesis on the glass surface by photolithography [28–30]. The strength of this approach lies in its ability to discriminate DNA molecules based on single base-pair difference. This allows the application of this method to the fields of medical diagnos-

tics, pharmacogenetics, and sequencing by hybridization as well as gene-expression analysis.

Fabrication of oligonucleotide chips by photolithography is theoretically simple but technically complex [29,30]. The light from a high-intensity mercury lamp is directed through a photolithographic mask onto the silica surface, resulting in deprotection of the terminal nucleotides in the illuminated regions. The entire chip is then reacted with the desired free nucleotide, resulting in selected chain elongation. This process requires only $4n$ cycles (where n = oligonucleotide length in bases) to synthesize a vast number of unique oligos, the total number of which is limited only by the complexity of the photolithographic mask and the chip size [29,31,32].

Sample preparation involves the generation of double-stranded cDNA from cellular poly(A)⁺ RNA followed by antisense RNA synthesis in an in vitro transcription reaction with biotinylated or fluor-tagged nucleotides. The RNA probe is then fragmented to facilitate hybridization. If the indirect visualization method is used, the chips are incubated with fluor-linked streptavidin (e.g., phycoerythrin) after hybridization [12,33]. The signal is detected with a custom confocal scanner [34]. This method has been applied successfully to the mapping of genomic library clones [35], to de novo sequencing by hybridization [28,36], and to evolutionary sequence comparison of the *BRCA1* gene [37]. In addition, mutations in the cystic fibrosis [38] and *BRCA1* [39] gene products and polymorphisms in the human immunodeficiency virus-1 clade B protease gene [40] have been detected by this method. Oligonucleotide chips are also useful for expression monitoring [33] as has been demonstrated by the simultaneous evaluation of gene-expression patterns in nearly all open reading frames of the yeast strain *S. cerevisiae* [12]. More recently, oligonucleotide chips have been used to help identify single nucleotide polymorphisms in the human [41] and yeast [42] genomes.

THE USE OF MICROARRAYS IN TOXICOLOGY

Screening for Mechanism of Action

The field of toxicology uses numerous in vivo model systems, including the rat, mouse, and rabbit, to assess potential toxicity and these bioassays are the mainstay of toxicology testing. However, in the past several decades, a plethora of in vitro techniques have been developed to measure toxicity, many of which measure toxicant-induced DNA damage. Examples of these assays include the Ames test, the Syrian hamster embryo cell transformation assay, micronucleus assays, measurements of sister chromatid exchange and unscheduled DNA synthesis, and many others. Fundamental to all of these methods is the fact that toxicity is often preceded by, and results in, alterations in gene expression. In many cases, these changes in gene expression are a

far more sensitive, characteristic, and measurable endpoint than the toxicity itself. We therefore propose that a method based on measurements of the genome-wide gene expression pattern of an organism after toxicant exposure is fundamentally informative and complements the established methods described above.

We are developing a method by which toxicants can be identified and their putative mechanisms of action determined by using toxicant-induced gene expression profiles. In this method, in one or more defined model systems, dose and time-course parameters are established for a series of toxicants within a given prototypic class (e.g., polycyclic aromatic hydrocarbons (PAHs)). Cells are then treated with these agents at a fixed toxicity level (as measured by cell survival), RNA is harvested, and toxicant-induced gene expression changes are assessed by hybridization to a cDNA microarray chip (Figure 1). We have developed a custom DNA chip, called ToxChip v1.0, specifically for this purpose and will discuss it in more detail below. The changes in gene expression induced by the test agents in the model systems are analyzed, and the common set of changes unique to that class of toxicants, termed a toxicant signature, is determined.

This signature is derived by ranking across all experiments the gene-expression data based on rela-

tive fold induction or suppression of genes in treated samples versus untreated controls and selecting the most consistently different signals across the sample set. A different signature may be established for each prototypic toxicant class. Once the signatures are determined, gene-expression profiles induced by unknown agents in these same model systems can then be compared with the established signatures. A match assigns a putative mechanism of action to the test compound. Figure 2 illustrates this signature method for different types of oxidant stressors, PAHs, and peroxisome proliferators. In this example, the unknown compound in question had a gene-expression profile similar to that of the oxidant stressors in the database. We anticipate that this general method will also reveal cross talk between different pathways induced by a single agent (e.g., reveal that a compound has both PAH-like and oxidant-like properties). In the future, it may be necessary to distinguish very subtle differences between compounds within a very large sample set (e.g., thousands of highly similar structural isomers in a combinatorial chemistry library or peptide library). To generate these highly refined signatures, standard statistical clustering techniques or principal-component analysis can be used.

For the studies outlined in Figure 2, we developed the custom cDNA microarray chip ToxChip v1.0.

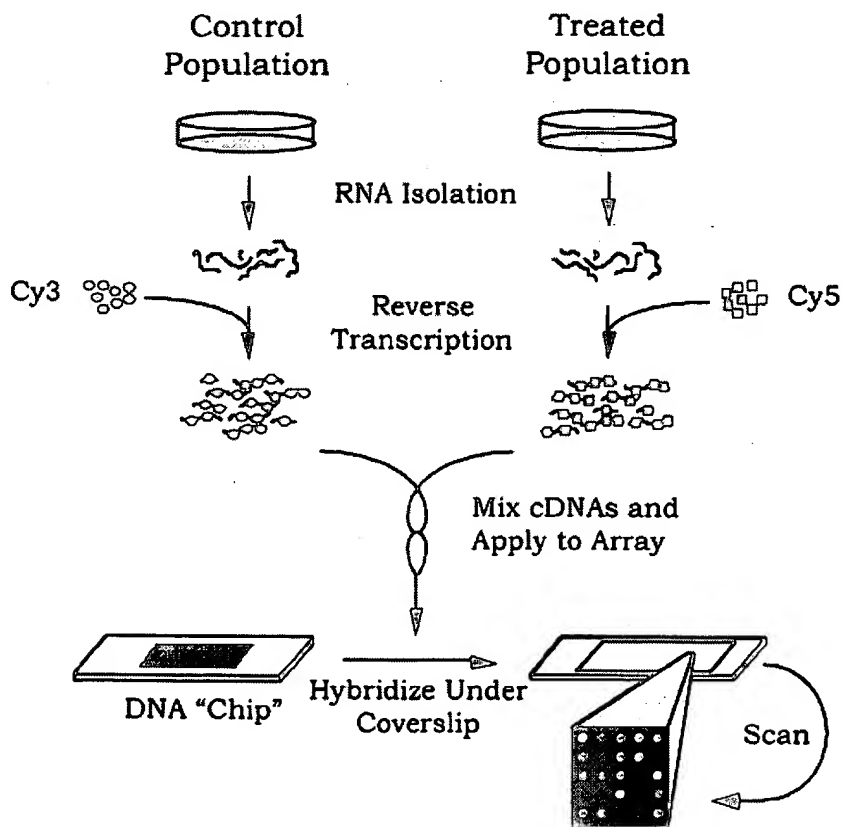


Figure 1. Simplified overview of the method for sample preparation and hybridization to cDNA microarrays. For illus-

trative purposes, samples derived from cell culture are depicted, although other sample types are amenable to this analysis.

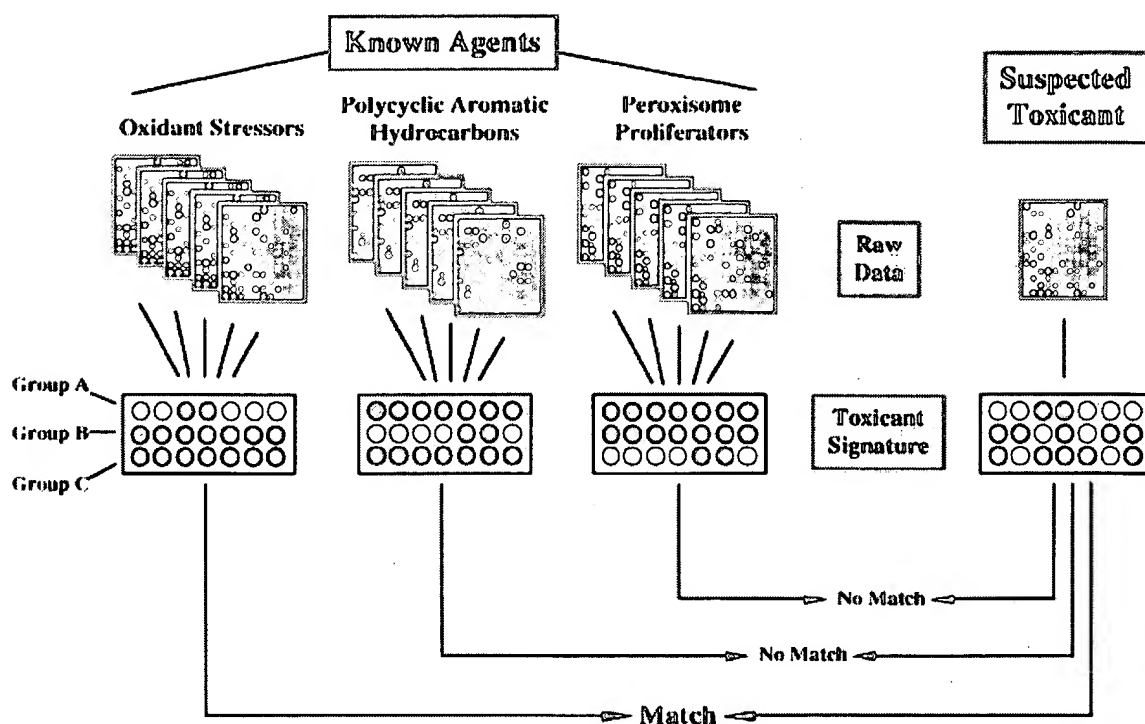


Figure 2. Schematic representation of the method for identification of a toxicant's mechanism of action. In this method, gene-expression data derived from exposure of model systems to known toxicants are analyzed, and a set of changes characteristic to that type of toxicant (termed the toxicant signature) is identified. As depicted, oxidant stressors produce

consistent changes in group A genes (indicated by red and green circles), but not group B or C genes (indicated by gray circles). The set of gene-expression changes elicited by the suspected toxicant is then compared with these characteristic patterns, and a putative mechanism of action is assigned to the unknown agent.

The 2090 human genes that comprise this subarray were selected for their well-documented involvement in basic cellular processes as well as their responses to different types of toxic insult. Included on this list are DNA replication and repair genes, apoptosis genes, and genes responsive to PAHs and dioxin-like compounds, peroxisome proliferators, estrogenic compounds, and oxidant stress. Some of the other categories of genes include transcription factors, oncogenes, tumor suppressor genes, cyclins, kinases, phosphatases, cell adhesion and motility genes, and homeobox genes. Also included in this group are 84 housekeeping genes, whose hybridization intensity is averaged and used for signal normalization of the other genes on the chip. To date, very few toxicants have been shown to have appreciable effects on the expression of these housekeeping genes. However, this housekeeping list will be revised if new data warrant the addition or deletion of a particular gene. Table 1 contains a general description of some of the different classes of genes that comprise ToxChip v1.0.

When a toxicant signature is determined, the genes within this signature are flagged within the database. When uncharacterized toxicants are then screened, the data can be quickly reformatted so that blocks of genes representing the different signatures

are displayed [11]. This facilitates rapid, visual interpretation of data. We are also developing ToxChip v2.0 and chips for other model systems, including rat, mouse, *Xenopus*, and yeast, for use in toxicology studies.

Animal Models in Toxicology Testing

The toxicology community relies heavily on the use of animals as model systems for toxicology testing. Unfortunately, these assays are inherently expensive, require large numbers of animals and take a long time to complete and analyze. Therefore, the National Institute of Environmental Health Sciences (NIEHS), the National Toxicology Program, and the toxicology community at large are committed to reducing the number of animals used, by developing more efficient and alternative testing methodologies. Although substantial progress has been made in the development of alternative methods, bioassays are still used for testing endpoints such as neurotoxicity, immunotoxicity, reproductive and developmental toxicology, and genetic toxicology. The rodent cancer bioassay is a particularly expensive and time-consuming assay, as it requires almost 4 yr, 1200 animals, and millions of dollars to execute and analyze [43]. In vitro experiments of the type outlined in Figure 2 might provide evidence that an unknown

Table 1. ToxChip v1.0: A Human cDNA Microarray Chip Designed to Detect Responses to Toxic Insult

Gene category	No. of genes on chip
Apoptosis	72
DNA replication and repair	99
Oxidative stress/redox homeostasis	90
Peroxisome proliferator responsive	22
Dioxin/PAH responsive	12
Estrogen responsive	63
Housekeeping	84
Oncogenes and tumor suppressor genes	76
Cell-cycle control	51
Transcription factors	131
Kinases	276
Phosphatases	88
Heat-shock proteins	23
Receptors	349
Cytochrome P450s	30

*This list is intended as a general guide. The gene categories are not unique, and some genes are listed in multiple categories.

agent is (or is not) responsible for eliciting a given biological response. This information would help to select a bioassay more specifically suited to the agent in question or perhaps suggest that a bioassay is not necessary, which would dramatically reduce cost, animal use, and time.

The addition of microarray techniques to standard bioassays may dramatically enhance the sensitivity and interpretability of the bioassay and possibly reduce its cost. Gene-expression signatures could be determined for various types of tissue-specific toxicants, and new compounds could be screened for these characteristic signatures, providing a rapid and sensitive *in vivo* test. Also, because gene expression is often exquisitely sensitive to low doses of a toxicant, the combination of gene-expression screening and the bioassay might allow the use of lower toxicant doses, which are more relevant to human exposure levels, and the use of fewer animals. In addition, gene-expression changes are normally measured in hours or days, not in the months to years required for tumor development. Furthermore, microarrays might be particularly useful for investigating the relationship between acute and chronic toxicity and identifying secondary effects of a given toxicant by studying the relationship between the duration of exposure to a toxicant and the gene-expression profile produced. Thus, a bioassay that incorporates gene-expression signatures with traditional endpoints might be substantially shorter, use more realistic dose regimens, and cost substantially less than the current assays do.

These considerations are also relevant for branches of toxicology not related to human health and not using rodents as model systems, such as aquatic toxicology and plant pathology. Bioassays based on the flathead minnow, *Daphnia*, and *Arabidopsis* could

also be improved by the addition of microarray analysis. The combination of microarrays with traditional bioassays might also be useful for investigating some of the more intractable problems in toxicology research, such as the effects of complex mixtures and the difficulties in cross-species extrapolation.

Exposure Assessment, Environmental Monitoring, and Drug Safety

The currently used methods for assessment of exposure to chemical toxicants are based on measurement of tissue toxin levels or on surrogate markers of toxicity, termed biomarkers (e.g., peripheral blood levels of hepatic enzymes or DNA adducts). Because gene expression is a sensitive endpoint, gene expression as measured with microarray technology may be useful as a new biomarker to more precisely identify hazards and to assess exposure. Similarly, microarrays could be used in an environmental-monitoring capacity to measure the effect of potential contaminants on the gene-expression profiles of resident organisms. In an analogous fashion, microarrays could be used to measure gene-expression endpoints in subjects in clinical trials. The combination of these gene-expression data and more established toxic endpoints in these trials could be used to define highly precise surrogates of safety.

Gene-expression profiles in samples from exposed individuals could be compared to the profiles of the same individuals before exposure. From this information, the nature of the toxic exposure can be determined or a relative clinical safety factor estimated. In the future it may also be possible to estimate not only the nature but the dose of the toxicant for a given exposure, based on relative gene-expression levels. This general approach may be particularly appropriate for occupational-health applications, in which unexposed and exposed samples from the same individuals may be obtainable. For example, a pilot study of gene expression in peripheral-blood lymphocytes of Polish coke-oven workers exposed to PAHs (and many other compounds) is under consideration at the NIEHS. An important consideration for these types of studies is that gene expression can be affected by numerous factors, including diet, health, and personal habits. To reduce the effects of these confounding factors, it may be necessary to compare pools of control samples with pools of treated samples. In the future it may be possible to compare exposed sample sets to a national database of human-expression data, thus eliminating the need to provide an unexposed sample from the same individual. Efforts to develop such a national gene-expression database are currently under way [44,45]. However, this national database approach will require a better understanding of genome-wide gene expression across the highly diverse human population and of the effects of environmental factors on this expression.

Alleles, Oligo Arrays, and Toxicogenetics

Gene sequences vary between individuals, and this variability can be a causative factor in human diseases of environmental origin [46,47]. A new area of toxicology, termed toxicogenetics, was recently developed to study the relationship between genetic variability and toxicant susceptibility. This field is not the subject of this discussion, but it is worthwhile to note that the ability of oligonucleotide arrays to discriminate DNA molecules based on single base-pair differences makes these arrays uniquely useful for this type of analysis. Recent reports demonstrated the feasibility of this approach [41,42]. The NIEHS has initiated the Environmental Genome Project to identify common sequence polymorphisms in 200 genes thought to be involved in environmental diseases [48]. In a pilot study on the feasibility of this application to the Environmental Genome Project, oligonucleotide arrays will be used to resequence 20 candidate genes. This toxicogenetic approach promises to dramatically improve our understanding of interindividual variability in disease susceptibility.

FUTURE PRIORITIES

There are many issues that must be addressed before the full potential of microarrays in toxicology research can be realized. Among these are model system selection, dose selection, and the temporal nature of gene expression. In other words, in which species, at what dose, and at what time do we look for toxicant-induced gene expression? If human samples are analyzed, how variable is global gene expression between individuals, before and after toxicant exposure? What are the effects of age, diet, and other factors on this expression? Experience, in the form of large data sets of toxicant exposures, will answer these questions.

One of the most pressing issues for array scientists is the construction of a national public database (linked to the existing public databases) to serve as a repository for gene-expression data. This relational database must be made available for public use, and researchers must be encouraged to submit their expression data so that others may view and query the information. Researchers at the National Institutes of Health have made laudable progress in developing the first generation of such a database [44,45]. In addition, improved statistical methods for gene clustering and pattern recognition are needed to analyze the data in such a public database.

The proliferation of different platforms and methods for microarray hybridizations will improve sample handling and data collection and analysis and reduce costs. However, the variety of microarray methods available will create problems of data compatibility between platforms. In addition, the near-infinite variety of experimental conditions under

which data will be collected by different laboratories will make large-scale data analysis extremely difficult. To help circumvent these future problems, a set of standards to be included on all platforms should be established. These standards would facilitate data entry into the national database and serve as reference points for cross-platform and inter-laboratory data analysis.

Many issues remain to be resolved, but it is clear that new molecular techniques such as microarray hybridization will have a dramatic impact on toxicology research. In the future, the information gathered from microarray-based hybridization experiments will form the basis for an improved method to assess the impact of chemicals on human and environmental health.

ACKNOWLEDGMENTS

The authors would like to thank Drs. Robert Maronpot, George Lucier, Scott Masten, Nigel Walker, Raymond Tennant, and Ms. Theodora Deverenux for critical review of this manuscript. EFN was supported in part by NIEHS Training Grant #ES07017-24.

REFERENCES

1. <http://www.ncbi.nlm.nih.gov/Web/Genbank/index.html>
2. <http://www.ncbi.nlm.nih.gov/Entrez/Genome/org.html>
3. Fleischmann RD, Adams MD, White O, et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 1995;269:496-512.
4. Goffeau A, Barrell BG, Bussey H, et al. Life with 6000 genes. *Science* 1996;274:546, 563-567.
5. <http://www.perkin-elmer.com/press/prc5448.html>
6. Liang P, Pardee AB. Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science* 1992;257:967-971.
7. Pietu G, Alibert O, Guichard V, et al. Novel gene transcripts preferentially expressed in human muscles revealed by quantitative hybridization of a high density cDNA array. *Genome Res* 1996;6:492-503.
8. Zhao ND, Hashida H, Takahashi N, Misumi Y, Sakaki Y. High-density cDNA filter analysis—A novel approach for large-scale, quantitative analysis of gene expression. *Gene* 1995;156:207-213.
9. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. *Science* 1995;270:484-487.
10. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene-expression patterns with a complementary DNA microarray. *Science* 1995;270:467-470.
11. DeRisi J, Penland L, Brown PO, et al. use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat Genet* 1996;14:457-460.
12. Wodicka L, Dong HL, Mittmann M, Ho MH, Lockhart DJ. Genome-wide expression monitoring in *Saccharomyces cerevisiae*. *Nat Biotechnol* 1997;15:1359-1367.
13. Marshall A, Hodgson J. DNA chips: An array of possibilities. *Nat Biotechnol* 1998;16:27-31.
14. <http://www.synteni.com>
15. Shalon D, Smith SJ, Brown PO. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res* 1996;6:639-645.
16. Chen Y, Dougherty ER, Bittner ML. Ratio-based decisions and the quantitative analysis of cDNA microarray images. *Biomedical Optics* 1997;2:364-374.
17. Khan J, Simon R, Bittner M, et al. Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays. *Cancer Res* 1998;58:5009-5013.
18. Schena M, Shalon D, Heller R, Chai A, Brown PO, Davis RW. Parallel human genome analysis: Microarray-based expression monitoring of 1000 genes. *Proc Natl Acad Sci USA* 1996; 93:10614-10619.

19. Lashkari DA, DeRisi JL, McCusker JH, et al. Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proc Natl Acad Sci USA* 1997;94:13057-13062.
20. Heller RA, Schena M, Chai A, et al. Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. *Proc Natl Acad Sci USA* 1997;94:2150-2155.
21. DeRisi JL, Iyer VR, Brown PO. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 1997;278:680-686.
22. Drmanac S, Stavropoulos NA, Labat I, et al. Gene-representing cDNA clusters defined by hybridization of 57,419 clones from infant brain libraries with short oligonucleotide probes. *Genomics* 1996;37:29-40.
23. Milosavljevic A, Savkovic S, Crkvenjakov R, et al. DNA sequence recognition by hybridization to short oligomers: Experimental verification of the method on the *E. coli* genome. *Genomics* 1996;37:77-86.
24. Drmanac S, Drmanac R. Processing of cDNA and genomic kilobase-size clones for massive screening, mapping and sequencing by hybridization. *Biotechniques* 1994;17:328-329, 332-336.
25. <http://www.resgen.com/>
26. <http://www.genomesystems.com/>
27. <http://www.clontech.com/>
28. Pease AC, Solas DA, Fodor SPA. Parallel synthesis of spatially addressable oligonucleotide probe matrices. *Abstracts of Papers of the American Chemical Society* 1992;203:34.
29. Pease AC, Solas D, Sullivan EJ, Cronin MT, Holmes CP, Fodor SPA. Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proc Natl Acad Sci USA* 1994;91:5022-5026.
30. Fodor SPA, Read JL, Pirrung MC, Stryer L, Lu AT, Solas D. Light-directed, spatially addressable parallel chemical synthesis. *Science* 1991;251:767-773.
31. McGall G, Labadie J, Brock P, Wallraff G, Nguyen T, Hinsberg W. Light-directed synthesis of high-density oligonucleotide arrays using semiconductor photoresists. *Proc Natl Acad Sci USA* 1996;93:13555-13560.
32. Lipshutz RJ, Morris D, Chee M, et al. Using oligonucleotide probe arrays to access genetic diversity. *Biotechniques* 1995;19:442-447.
33. Lockhart DJ, Dong HL, Byrne MC, et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol* 1996;14:1675-1680.
34. <http://www.mdyn.com/>
35. Sapolsky RJ, Lipshutz RJ. Mapping genomic library clones using oligonucleotide arrays. *Genomics* 1996;33:445-456.
36. Chee M, Yang R, Hubbell E, et al. Accessing genetic information with high-density DNA arrays. *Science* 1996;274:610-614.
37. Hacia JG, Makalowski W, Edgemon K, et al. Evolutionary sequence comparisons using high-density oligonucleotide arrays. *Nat Genet* 1998;18:155-158.
38. Cronin MT, Fucini RV, Kim SM, Masino RS, Wespi RM, Miyada CG. Cystic fibrosis mutation detection by hybridization to light-generated DNA probe arrays. *Hum Mutat* 1996;7:244-255.
39. Hacia JG, Brody LC, Chee MS, Fodor SPA, Collins FS. Detection of heterozygous mutations in BRCA1 using high density oligonucleotide arrays and two-colour fluorescence analysis. *Nat Genet* 1996;14:441-447.
40. Kozal MJ, Shah N, Shen NP, et al. Extensive polymorphisms observed in HIV-1 clade B protease gene using high-density oligonucleotide arrays. *Nat Med* 1996;2:753-759.
41. Wang DG, Fan JB, Siao CJ, et al. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* 1998;280:1077-1082.
42. Winzeler EA, Richards DR, Conway AR, et al. Direct allelic variation scanning of the yeast genome. *Science* 1998;281:1194-1197.
43. Chhabra RS, Huff JE, Schwetz BS, Selkirk J. An overview of prechronic and chronic toxicity carcinogenicity experimental-study designs and criteria used by the National Toxicology Program. *Environ Health Perspect* 1990;86:313-321.
44. Ermolaeva O, Rastogi M, Pruitt KD, et al. Data management and analysis for gene expression arrays. *Nat Genet* 1998;20:19-23.
45. <http://www.nhgri.nih.gov/DIR/LCG/15K/HTML/dbase.html>
46. Samson M, Libert F, Doranz BJ, et al. Resistance to HIV-1 infection in Caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene. *Nature* 1996;382:722-725.
47. Bell DA, Taylor JA, Paulson DF, Robertson CN, Mohler JL, Lucier GW. Genetic risk and carcinogen exposure—A common inherited defect of the carcinogen-metabolism gene glutathione-S-transferase M1 (*Gstm1*) that increases susceptibility to bladder cancer. *J Natl Cancer Inst* 1993;85:1159-1164.
48. <http://www.niehs.nih.gov/envgenom/home.html>



Expression profiling in toxicology — potentials and limitations

Sandra Steiner *, N. Leigh Anderson

Large Scale Biology Corporation, 9620 Medical Center Drive, Rockville, MD 20850-3338, USA

Abstract

Recent progress in genomics and proteomics technologies has created a unique opportunity to significantly impact the pharmaceutical drug development processes. The perception that cells and whole organisms express specific inducible responses to stimuli such as drug treatment implies that unique expression patterns, molecular fingerprints, indicative of a drug's efficacy and potential toxicity are accessible. The integration into state-of-the-art toxicology of assays allowing one to profile treatment-related changes in gene expression patterns promises new insights into mechanisms of drug action and toxicity. The benefits will be improved lead selection, and optimized monitoring of drug efficacy and safety in pre-clinical and clinical studies based on biologically relevant tissue and surrogate markers. © 2000 Elsevier Science Ireland Ltd. All rights reserved.

Keywords: Proteomics; Genomics; Toxicology

1. Introduction

The majority of drugs act by binding to protein targets, most to known proteins representing enzymes, receptors and channels, resulting in effects such as enzyme inhibition and impairment of signal transduction. The treatment-induced perturbations provoke feedback reactions aiming to compensate for the stimulus, which almost always are associated with signals to the nucleus, resulting in altered gene expression. Such gene expression regulations account for both the

pharmacological action and the toxicity of a drug and can be visualized by either global mRNA or global protein expression profiling. Hence, for each individual drug, a characteristic gene regulation pattern, its molecular fingerprint, exists which bears valuable information on its mode of action and its mechanism of toxicity.

Gene expression is a multistep process that results in an active protein (Fig. 1). There exist numerous regulation systems that exert control at and after the transcription and the translation step. Genomics, by definition, encompasses the quantitative analysis of transcripts at the mRNA level, while the aim of proteomics is to quantify gene expression further down-stream, creating a snapshot of gene regulation closer to ultimate cell function control.

* Corresponding author. Tel.: +1-301-4245989; fax: +1-301-7624892.

E-mail address: steiner@lsbc.com (S. Steiner)

2. Global mRNA profiling

Expression data at the mRNA level can be produced using a set of different technologies such as DNA microarrays, reverse transcript imaging, amplified fragment length polymorphism (AFLP), serial analysis of gene expression (SAGE) and others. Currently, DNA microarrays are very popular and promise a great potential. On a typical array, each gene of interest is represented either by a long DNA fragment (200–2400 bp) typically generated by polymerase chain reaction (PCR) and spotted on a suitable substrate using robotics (Schena et al., 1995; Shalon et al., 1996) or by several short oligonucleotides (20–30 bp) synthesized directly onto a solid support using photolabile nucleotide chemistry (Fodor et al., 1991; Chee et al., 1996). From control and treated tissues, total RNA or mRNA is isolated and reverse transcribed in the presence of radioactive or fluorescent labeled nucleotides, and the labeled probes are then hybridized to the arrays. The intensity of the array signal is measured for each gene transcript by either autoradiography or laser scanning confocal microscopy. The ratio between the signals of control and treated samples reflect the relative drug-induced change in transcript abundance.

3. Global protein profiling

Global quantitative expression analysis at the protein level is currently restricted to the use of two-dimensional gel electrophoresis. This technique combines separation of tissue proteins by isoelectric focusing in the first dimension and by sodium dodecyl sulfate slab gel electrophoresis-based molecular weight separation on the second, orthogonal dimension (Anderson et al., 1991). The product is a rectangular pattern of protein spots that are typically revealed by Coomassie Blue, silver or fluorescent staining (Fig. 2). Protein spots are identified by mass spectrometry following generation of peptide mass fingerprints (Mann et al., 1993) and sequence tags (Wilkins et al., 1996). Similar to the mRNA approach, the ratio between the optical density of spots from control and treated samples are compared to search for treatment-related changes.

4. Expression data analysis

Bioinformatics forms a key element required to organize, analyze and store expression data from either source, the mRNA or the protein level. The overall objective, once a mass of high-quality

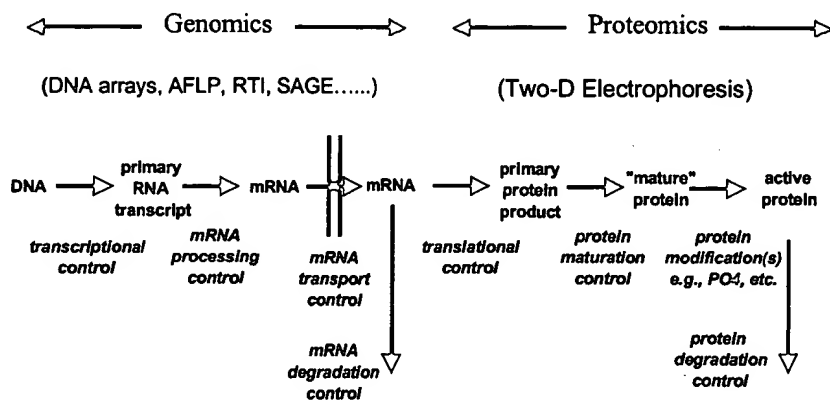


Fig. 1. Production of an active protein is a multistep process in which numerous regulation systems exert control at various stages of expression. Molecular fingerprints of drugs can be visualized through expression profiling at the mRNA level (genomics) using a variety of technologies and at the protein level (proteomics) using two-dimensional gel electrophoresis.

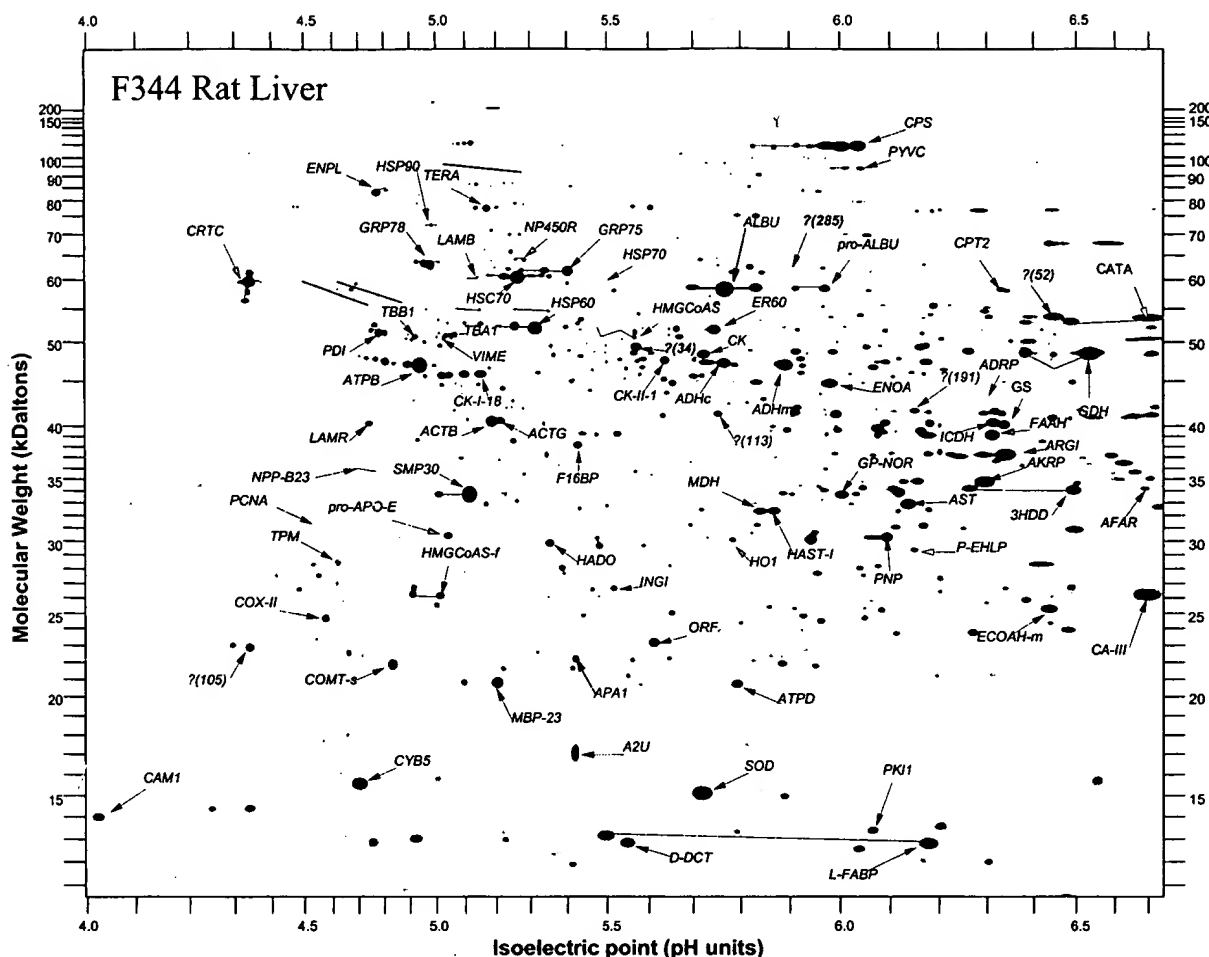


Fig. 2. Computerized representation of a Coomassie Blue stained two-dimensional gel electrophoresis pattern of Fischer F344 rat liver homogenate.

quantitative expression data has been collected, is to visualize complex patterns of gene expression changes, to detect pathways and sets of genes tightly correlated with treatment efficacy and toxicity, and to compare the effects of different sets of treatment (Anderson et al., 1996). As the drug effect database is growing, one may detect similarities and differences between the molecular fingerprints produced by various drugs, information that may be crucial to make a decision whether to refocus or extend the therapeutic spectrum of a drug candidate.

5. Comparison of global mRNA and protein expression profiling

There are several synergies and overlaps of data obtained by mRNA and protein expression analysis. Low abundant transcripts may not be easily quantified at the protein level using standard two-dimensional gel electrophoresis analysis and their detection may require prefractionation of samples. The expression of such genes may be preferably quantified at the mRNA level using techniques allowing PCR-mediated target ampli-

cation. Tissue biopsy samples typically yield good quality of both mRNA and proteins; however, the quality of mRNA isolated from body fluids is often poor due to the faster degradation of mRNA when compared with proteins. RNA samples from body fluids such as serum or urine are often not very 'meaningful', and secreted proteins are likely more reliable surrogate markers for treatment efficacy and safety. Detection of post-translational modifications, events often related to function or nonfunction of a protein, is restricted to protein expression analysis and rarely can be predicted by mRNA profiling. Information on subcellular localization and translocation of proteins has to be acquired at the level of the protein in combination with sample prefractionation procedures. The growing evidence of a poor correlation between mRNA and protein abundance (Anderson and Seilhamer, 1997) further suggests that the two approaches, mRNA and protein profiling, are complementary and should be applied in parallel.

6. Expression profiling and drug development

Understanding the mechanisms of action and toxicity, and being able to monitor treatment efficacy and safety during trials is crucial for the successful development of a drug. Mechanistic insights are essential for the interpretation of drug effects and enhance the chances of recognizing potential species specificities contributing to an improved risk profile in humans (Richardson et al., 1993; Steiner et al., 1996b; Aicher et al., 1998). The value of expression profiling further increases when links between treatment-induced expression profiles and specific pharmacological and toxic endpoints are established (Anderson et al., 1991, 1995, 1996; Steiner et al. 1996a). Changes in gene expression are known to precede the manifestation of morphological alterations, giving expression profiling a great potential for early compound screening, enabling one to select drug candidates with wide therapeutic windows reflected by molecular fingerprints indicative of high pharmacological potency and low toxicity (Arce et al., 1998). In later phases of drug devel-

opment, surrogate markers of treatment efficacy and toxicity can be applied to optimize the monitoring of pre-clinical and clinical studies (Doherty et al., 1998).

7. Perspectives

The basic methodology of safety evaluation has changed little during the past decades. Toxicity in laboratory animals has been evaluated primarily by using hematological, clinical chemistry and histological parameters as indicators of organ damage. The rapid progress in genomics and proteomics technologies creates a unique opportunity to dramatically improve the predictive power of safety assessment and to accelerate the drug development process. Application of gene and protein expression profiling promises to improve lead selection, resulting in the development of drug candidates with higher efficacy and lower toxicity. The identification of biologically relevant surrogate markers correlated with treatment efficacy and safety bears a great potential to optimize the monitoring of pre-clinical and clinical trials.

References

- Aicher, L., Wahl, D., Arce, A., Grenet, O., Steiner, S., 1998. New insights into cyclosporine A nephrotoxicity by proteome analysis. *Electrophoresis* 19, 1998–2003.
- Anderson, N.L., Seilhamer, J., 1997. A comparison of selected mRNA and protein abundances in human liver. *Electrophoresis* 18, 533–537.
- Anderson, N.L., Esquer-Blasco, R., Hofmann, J.P., Anderson, N.G., 1991. A two-dimensional gel database of rat liver proteins useful in gene regulation and drug effects studies. *Electrophoresis* 12, 907–930.
- Anderson, L., Steele, V.K., Kelloff, G.J., Sharma, S., 1995. Effects of oltipraz and related chemoprevention compounds on gene expression in rat liver. *J. Cell. Biochem. Suppl.* 22, 108–116.
- Anderson, N.L., Esquer-Blasco, R., Richardson, F., Foxworthy, P., Eacho, P., 1996. The effects of peroxisome proliferators on protein abundances in mouse liver. *Toxicol. Appl. Pharmacol.* 137, 75–89.
- Arce, A., Aicher, L., Wahl, D., Esquer-Blasco, R., Anderson, N.L., Cordier, A., Steiner, S., 1998. Changes in the liver proteome of female Wistar rats treated with the hypoglycemic agent SDZ PGU 693. *Life Sci.* 63, 2243–2250.

- Chee, M., Yang, R., Hubbell, E., Berno, A., Huang, X.C., Stern, D., Winkler, J., Lockhart, D.J., Morris, M.S., Fodor, S.P., 1996. Accessing genetic information with high-density DNA arrays. *Science* 274, 610–614.
- Doherty, N.S., Littman, B.H., Reilly, K., Swindell, A.C., Buss, J., Anderson, N.L., 1998. Analysis of changes in acute-phase plasma proteins in an acute inflammatory response and in rheumatoid arthritis using two-dimensional gel electrophoresis. *Electrophoresis* 19, 355–363.
- Fodor, S.P., Read, J.L., Pirrung, M.C., Stryer, L., Lu, A.T., Solas, D., 1991. Light-directed, spatially addressable parallel chemical synthesis. *Science* 251, 767–773.
- Mann, M., Hojrup, P., Roepstorff, P., 1993. Use of mass spectrometric molecular weight information to identify proteins in sequence databases. *Biol. Mass Spectrom.* 22, 338–345.
- Richardson, F.C., Strom, S.C., Copple, D.M., Bendele, R.A., Probst, G.S., Anderson, N.L., 1993. Comparisons of protein changes in human and rodent hepatocytes induced by the rat-specific carcinogen, methapyrilene. *Electrophoresis* 14, 157–161.
- Schena, M., Shalon, D., Davis, R.W., Brown, P.O., 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 251, 467–470.
- Shalon, D., Smith, S.J., Brown, P.O., 1996. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res.* 6, 639–645.
- Steiner, S., Wahl, D., Mangold, B.L.K., Robison, R., Raynackers, J., Meheus, L., Anderson, N.L., Cordier, A., 1996a. Induction of the adipose differentiation-related protein in liver of etomoxir treated rats. *Biochem. Biophys. Res. Commun.* 218, 777–782.
- Steiner, S., Aicher, L., Raymackers, J., Meheus, L., Esquer-Blasco, R., Anderson, L., Cordier, A., 1996b. Cyclosporine A mediated decrease in the rat renal calcium binding protein calbindin-D 28 kDa. *Biochem. Pharmacol.* 51, 253–258.
- Wilkins, M.R., Gasteiger, E., Sanchez, J.C., Appel, R.D., Hochstrasser, D.F., 1996. Protein identification with sequence tags. *Curr. Biol.* 6, 1543–1544.

Subject: RE: [Fwd: Toxicology Chip]**Date: Mon. 3 Jul 2000 08:09:45 -0400****From: "Afshari, Cynthia" <afshari@niehs.nih.gov>****To: "Diana Hamlet-Cox" <dianahc@incyte.com>**

You can see the list of clones that we have on our 12K chip at

<http://manuel.niehs.nih.gov/maps/guest/clonesrch.cfm>

We selected a subset of genes (2000K) that we believed critical to tox response and basic cellular processes and added a set of clones and ESTs to this. We have included a set of control genes (80+) that were selected by the NHGRI because they did not change across a large set of array experiments. However, we have found that some of these genes change significantly after tox treatments and are in the process of looking at the variation of each of these 80+ genes across our experiments.

Our chips are constantly changing and being updated and we hope that our data will lead us to what the toxchip should really be.

I hope this answers your question.

Cindy Afshari

> -----

> From: Diana Hamlet-Cox
> Sent: Monday, June 26, 2000 8:52 PM
> To: afshari@niehs.nih.gov
> Subject: [Fwd: Toxicology Chip]

> Dear Dr. Afshari,

> Since I have not yet had a response from Bill Grigg, perhaps he was not
> the right person to contact.

> Can you help me in this matter? I don't need to know the sequences,
> necessarily, but I would like very much to know what types of sequences
> are being used, e.g., GPCRs (more specific?), ion channels, etc.

> Diana Hamlet-Cox

> ----- Original Message -----

> Subject: Toxicology Chip
> Date: Mon, 19 Jun 2000 18:31:48 -0700
> From: Diana Hamlet-Cox <dianahc@incyte.com>
> Organization: Incyte Pharmaceuticals
> To: grigg@niehs.nih.gov

> Dear Colleague:

> I am doing literature research on the use of expressed genes as
> pharmacotoxicology markers, and found the Press Release dated February
> 29, 2000 regarding the work of the NIEHS in this area. I would like to
> know if there is a resource I can access (or you could provide?) that
> would give me a list of the 12,000 genes that are on your Human ToxChip
> Microarray. In particular, I am interested in the criteria used to
> select sequences for the ToxChip, including any control sequences
> included in the microarray.

> Thank you for your assistance in this request.

> Diana Hamlet-Cox, Ph.D.
> Incyte Genomics, Inc.

> --

> =====

> This email message is for the sole use of the intended recipient s and
> may contain confidential and privileged information subject to
> attorney-client privilege. Any unauthorized review, use, disclosure or
> distribution is prohibited. If you are not the intended recipient,
> please contact the sender by reply email and destroy all copies of the
> original message.

> =====

>
>
>